378.744
BO
AM 1932
je

Boston University
College of Liberal Arts
Library

THE GIFT OF the Author

378.744
BO
AM1932

48510 je

BOSTON UNIVERSITY

GRADUATE SCHOOL

THE RATING SCALE AND RELIGIOUS EDUCATION

A Thesis

Submitted by

Walter L. Jenkins

(B.R.E., Boston University, 1928)

In partial fulfillment of requirements

for the degree of Master of Arts.

1932

BOSTON UNIVERSITY
COLLEGE OF LIBERAL ARTS
LIBRARY

48510

378.744
B0
A.M.1932
je

TABLE OF CONTENTS

I Introduction - The Need for Measurement in Religious Education.

A. The Need for Refined Measurement.....1

B. Definition of Religious Education.....7
in Terms of Objectives.

C. What Ought We to Try to Measure in Religious Education?.....10

II Significant Developments in the Field of Measurement.

A. Historical Sketch of the Measurement Movement. 1

B. Measurement in the Field of Religious Education at Present. .....

III The Rating Scale

A. The Theory of the Rating Scale. .....

B. Types of Rating Scales. 31

C. Criticism of Rating Scales. .....

IV The Rating Scale in Religious Education

A. The Rating Scale - A Supervisory Instrument. .....

B. The Rating of Teachers. .....

C. The Rating of Pupils. 101

D. Rating the Program .....

V Conclusion.....

VI Summary.....117

Bibliography.....

Chapter I

The Need for Measurement in Religious Education

Digitized by the Internet Archive
in 2016

https://archive.org/details/theratingscalere00jenk

. We are making choices based on some sort of evaluation every day of our lives. Comparisons, either qualitative or quantitative are the basis of measurement, whether they are stated or implied. To state or to assume that Mr. Jones is cultured or well educated is common practice, and yet we would hardly dignify it by calling it measurement. The question is <u>how well</u> educated or <u>how</u> cultured is Mr. Jones and on what do we base our judgment?

Watson (1) tells a story of a group of school men, who in the early days of measurement emphasis came together to talk about tests. One was fervently opposing the notion that certain aspects of mental life could be measured by any foot rule. He closed his protest with, "Who would presume to measure the intellect of a Milton or a Shakespeare?"

(1) Watson, G. B. "Experimentation and measurement in Religious Education" Association Press 1927, P.34.

Thorndike arose and answered, "Fortunately it is not necessary for us to measure the intellect of a Milton or a Shakespeare. That has already been done by the previous speaker. They have not only been measured but placed at the head of the list. It now remains only to find out where the rest of the human race stands with reference to them."

The politician who writes a letter of recommendation for a "friend" has as his purpose evaluation, whether it is real or imagined. This is in a sense measurement, but if he were to compare the one for whom he was writing with another individual, known to both the recipient and the writer of the letter, we could safely regard this as the first step in the refinement process.

There is great need for a refined, reliable system of measurement for all walks of life. The fact that a prospective army officer may be perfect in army routine and not have the faculty or leadership ability to command or lead men; that a prospective salesman may have complete knowledge of the material he is to handle, and not be able to make a sale; that a teacher may have every academic requirement and not be a good teacher; that a student may have knowledge of a body of information and lack the proper attitudes and abilities to put it to use;

that a person may subscribe to any system of dogma or creed and still not have any religion; that a college graduate may have the highest possible academic standing and still fail in one position after another, points immediately to the fact that there is an education outside the regularly accepted curriculum, and a sphere of experience which we cannot afford to overlook.

This phase of human make-up has been characterized by various terms, among which are dynamic qualities, character traits, and personality (in the broadest sense). The need for a system of measurement to indicate not only the presence or absence of t ese qualities but the degree of their presence or absence has been keenly felt since before 1910.

What type of man must I have for this position or that position?; or which one of these faithful employees ought I to make foreman?; are questions often asked in the field of business and industry. The principle of mass production and the necessity for economy demands a wise selection of personel. Professor Walter Dill Scott(1) and a seminar at Carnegie Institute of Technology, of which he was leader, created the first man-to-man rating scale, in an effort to meet this need.

(1) Rugg, Harold "Is the Rating of Human Character Practicable." Journal Educational Psychology, Nov.1921 P. 427.

Considerable progress has been made since then, but the need for better, more refined procedure is everywhere evident.

The World War found us in 1917 with a huge army to raise and equip and a host of officers to train. On what basis were men in the various Officer's Training Camps to be assigned to the several branches of the service? Who were the best and who were in line for promotion to important posts? Which ones were able to carry the most responsibility? The need for some scheme of evaluating the capacities and qualities of men was keenly felt. Sacrifice of time, material and even human life was the penalty for mistakes. The Army Rating Scale (1) was developed by Rugg and several others, under most ideal conditions and, was in a measure successful. Many additional experiments have been and are constantly being made, which is a very real testimony to the need of adequate measurement of abilities, capacities, habits and qualities.

In the field of Education, Boyce (2) points out very definite needs for measurement as it pertains to the

(1) Rugg, Harold-Journal of Education Psychology, Nov. 1921, Page 428 ("Is the Rating of Human Character Practicable")

(2) Boyce, Arthur Clifton-14th Yearbook, National Society for the Study of Education Part 2, PP 9-10

administration of a school or of a system of schools. The need exists first in the vocational guidance of teachers, particularly in aiding to discover the school grade and the type of work for which one is best fitted. Secondly, in the improvement of teachers in service by providing a basis for self-criticism and self-improvement. Thirdly, as a basis for the determination of promotion and dismissal, and salary schedule.

Another evidence of the need for measurement in the field of education is the inadequacy of the present basis of measurement, or system of grading. Robertson (1) points out that, "Education is a process of bringing about desired changes in people, and the present system is attempting to measure change in terms of A, B, C, E D, or O to 100 in knowledge of subjects." Thus only a portion of the total education of a person is being measured and that portion very inadequately. (2) Elliott and Starck showed that in a fairly objective subject like mathematics one hundred experienced teachers of the subject assigned, on the same set of actual replies to an examination paper, grades varying from 28 to 90.

(1) Robertson, David Allan - "Character Processes in Colleges and Universities." Religious Education, May, 1930 P.393

(2) " " " " " " " " " " " " " " " " "

The probability is that most objective examinations will not show as wide a degree of variance as the above, but the fact still remains that there is a degree of subjectivity in the most objective type of test.

Meanwhile, what of Religious Education? Very little has been done, but the need is perhaps more apparent here than in any other field. The great mass of teachers have little or no educational consciousness and ministers many times have less than the teachers. The spirit seems to be, - "We don't know where we're going but we're on our way." About the only widely used type of measurement, both in the so-called "Sunday School" and in the Church is attendance records as an indication of the quality of the program. Sometimes, the attendance records indicate something about program but they may also indicate the influence of high-pressure contests and kindred activities, which are a direct confession of lack of program.

The main problem facing the Church today is the creation of a trained leadership. How shall we know when this leadership is trained? When it has mastered a certain body of knowledge? Who is fitted to teach and who is not? What about the religious and social attitudes of boys and girls? How can we find out the presence or absence of religious, moral, social, ethical

ideals. A great need is apparent to even the casual observer.

B. DEFINITION OF RELIGIOUS EDUCATION IN TERMS OF OBJECTIVES.

What is Religious Education? Coe (1) says, "It is the systematic, critical examination and reconstruction of relations between persons guided by Jesus' assumption that persons are of infinite worth, and by the hypothesis of the existence of God, the Great Valuer of Persons". Many people would not be satisfied with this as a definition. However, we are not as much concerned with an academic definition of religious education as we are with some knowledge of the outcomes or goals which we strive to attain through the process of Religious Education.

Discussing the need for comprehensive objectives Bower (2) says, "In the light of our present knowledge of the spiritual needs of persons and society, the statement of general objectives might well assume four forms; in terms of personal life, the development of a complete, satisfying and effective Christian personality; in terms of knowledge, such acquaintance with racial religious experience as will help the learner to arrive

(1) Coe, George A.-"What is Christian Education?" P.296 Scribners 1929

(2) Bower, W.C.-"Religious Education in the Modern Church" Bethany Press 1929. P.36

at convictions of his own concerning the religious values of life; in terms of the Christian institution, an aware and effective Church as a specialized agency for the interpretation and promotion of Christian ideals and purposes; in terms of the great society, the gradual and progressive reconstruction of social relations and functions on a spiritual basis."

These four aims are indeed comprehensive. They appear individually as the dominant emphasis of the Church for given periods in its history. Fiske (1) describes the aims of Clement and Origen as the teaching of Christian virtues and the Christian education of body, mind and spirit (personal life); the aim for a thousand years after Constantine as keeping the Church alive and training leaders for it (an aware and effective Church); the aim of the reformation as a doctrinal or knowledge aim (knowledge); the aim of the early nineteenth century as evangelistic (the great society).

The most recent and widely accepted list of objectives for Religious Education is submitted by Veith (2).

(1) Fiske, G.W.-"Purpose in Teaching Religion" PP 42-45 Abbington Press 1927

(2) Veith, Paul H.-"Objectives in Religious Education" PP 80-89 Harpers 1930

1. "To foster in growing persons a consciousness of God as a reality in human experience, and a sense of personal relationship to him".

2. "To lead growing persons to an understanding and appreciation of the personality, life, and teachings of Jesus Christ".

3. "To foster in growing persons a progressive and continuous development of Christ-like character".

4. "To develop in growing persons the ability and disposition to participate in and contribute constructively to the building of a social order embodying the ideal of the fatherhood of God and the brotherhood of man".

5. "To lead growing persons to build a life philosophy on the basis of a Christian interpretation of life and the universe".

6. "To develop in growing persons the ability and disposition to participate in the organized society of Christians - the Church".

7. "To effect in growing persons the assimilation of the best religious experience of the race, as effective guidance to present experience".

No attempt is here made to make these seven objectives more specific. They comprise, as they are, a good working definition of religious education.

The assumption inherent in the above seven statements is that religious education is a process. The primary concern is persons and the progressive development of persons in their natural and normal relationships. The starting point is experience and the process is the understanding, analysis, criticism and evaluation of experience. Bower(1) says of the curriculum, "When it is approached in this way the curriculum of religious education becomes the experience of the learner as that experience undergoes interpretation, enrichment, and control in terms of religious ideas, ideals, and purposes". Religious Education is thus a voyage of discovery for each person rather than the assimilation of a body of material, preconceived ideas, etc.

Thus construed, religion becomes a quality of all life, and measurement in Religious Education becomes a technique of discovering the presence or absence and the degree of the presence or absence of the desirable qualities or objectives in the normal, every-day experience of the learner.

C. WHAT OUGHT WE TRY TO MEASURE IN RELIGIOUS EDUCATION.

(1) Lotz and Crawford-"Studies in Religious Education" P.182
Cokesbury Press 1931

The personalist would say we ought to attempt to measure Christian personality. This furnishes difficulties, however, in the possibility of misunderstanding in the use of the term. We usually think of personality as being the composite of tact, physical appearance, enthusiasm, etc., while character has a different connotation. McDougall defines character as follows: (1) "Character is the system of directed conative tendencies. It may be relatively simple or complex; it may be harmoniously organized or lacking in harmony, it may be firmly or loosely knit; it may be directed in the main toward lower or higher goals."

Two elements are evident in the above definition. First, active tendencies are directed. Secondly, active tendencies may be systematized or organized about dominant life purposes. Thus, character, as we popularly know it, has a moral connotation which we do not associate with the popular concept of personality.

Brandenburg (2) attempted to measure personality in its complete form, defining as accurately as possible his terms. He understood personality to indicate a composite of an individual's typical reactions, physical intellectual and emotional, to his environment, together with his various physical characteristics which constitute appearance".

(1) McDougall, Wm.--"Outline of Psychology--Scribners 1923 p.417.

(2) Brandenburg, George C.--"Analyzing Personality" Journal Applied Psychology June 1925 PP 139-155 Sept.1925 Pp 281-292.

Thus personality is construed as a broad term inclusive of what we popularly know as character. He used twenty-three traits in which twenty-nine students rated one another. The list of twenty-three traits is as follows:-PP 142-143:

1. Accuracy in Work.
2. Enthusiasm.
3. Aggressiveness.
4. Self-reliance.
5. Memory.
6. Popularity.
7. Motor Ability.
8. Tact.
9. General Ability.
10. Reliability.
11. Co-operation.
12. Reasoning Ability.
13. General Information.
14. Originality.
15. Sympathy.
16. Speed in Work.
17. Social & Civic interest.
18. Address.
19. Sincerity.
20. Industry.
21. Neatness.
22. Appreciation of humor.
23. Moral habits.

He developed some interesting inter-correlations, and some interesting conclusions regarding vocational life.

Unfortunately, this type of approach will not suffice for religious education. It is too general and not all-inclusive.

In direct contrast with this rather general attempt to measure "personality" Hartshorne and May (1) attempted a specific measurement of character and divided the field of

(1) Hartshorne, Hugh and May, Mark A "Testing the Knowledge of right and wrong" Monograph 1927

character study, in which tests are called for, as follows:

1. Mental content and skills, the so-called intellectual factors.

2. Desires, attitudes, motives, etc., the dynamic factors.

3. Social behavior, the performance factors.

4. Self-control, the relation of all these factors to one another and to social self-organization.

Thirteen tests were constructed in an effort to measure item one, and used in sufficient numbers to warrant statistical treatment. They were divided as follows:

A. Word Tests

1. Opposites - A multiple choice test in which the subject was required to write in the bracket the word most nearly opposite in meaning to the one in capital letters to the left.

1. GIVE 1-Present, 2-Accept, 3-take, 4-wish, 5-absent,
( ).

2. FRIEND 1-soldier, 2-true, 3-false, 4-enemy, 5-fight,
( ).

2. Similarities - A cross out test in which the subject was asked to cross out the odd word.

1. 1-debase, 2-ignore, 3-humble, 4-disgrace, 5-lower

2. 1-quit, 2-surrender, 3-enemy, 4-relinquish, 5-forsake

3. Word Consequences - A multiple choice test in which subject was asked to indicate: (1) likely consequences to actions indicated in capital letters, (2) the most likely consequences, (3) the best consequences, (4) the worst consequences.

1. CHEATING - 1-courage, 2-forgery, 3-outcast, 4-wealth, 5-poverty.

2. BETTING - 1-gambling, 2-poverty, 3-optimism, 4-wealth, 5-war.

B - Sentence Tests

4. Cause and effect - a true-false test with 100 items such as the following in which subject underlines true or false.

1. Good marks are chiefly a matter of luck - true false

2. Success always comes from hard work - true false

5. Duties - modified true - false test with modified response indicating whether the act is his duty, not his duty, sometimes, sometimes not his duty.

1. To help a slow or dull child with his lessons.
true ? false

2. To read the newspaper everyday. - true ? false

6. Comprehensions - multiple choice response to situations.

1. If someone asks to borrow your pencil

(a) Tell him it is broken.

(b) Tell him that you lost it.

(c) Tell him that you don't want to lend it.

(d) Let him take it.

7. Provocations - Illustrations. Stories of Children. Subject asked to decide whether right, wrong or excusable by encircling one of the possible answers.

1. Helen noticed that everyone in the class was cheating so she cheated too. ---R. Ex. Wr.

8. Foresights - No suggestion given to possible consequences. Subjects asked to fill in.

1. Whenever anyone picked on John, he would go and tell his teacher.

(Space for a large number of consequences).

9. Recognitions - Multiple choice test in which subject encircles response. C. For cheating, L. for Lying, S. for stealing, something wrong but neither cheating, lying or stealing encircle x, if not wrong encircle J.

1. Bullying younger children C L S X J
2. Using street car transfers which are out of date --- C L S X J

10. Principles - A true-false test

1. To master one's self is a greater thing than to win a battle _____ true, false.
2. Clean speech is a sign of being goody-goody _____ true, false.

11. Applications - A multiple choice test using elements in provocations and principle tests.

12. Social-ethical vocabulary - Subject places the number of the word, that means the same or most nearly the same as the word in Capitals, on the right.

1. BRAVERY - 1-folly, 2-courage, 3-livery, 4-Impertinence, 5-humanity _____.

2. SCOLD - 1-scold, 2-angry, 3-make fun of, 4-extol, 5-expound _____.

C. Good Manners Test - true, false.

A. - If soup or any liquid is too hot, blow on it slightly to cool it _____ true, false.

After preliminary experimentation these tests were revised and some of the tests were thrown out on the following bases:

1. Items with ambiguous or localized answers.

2. Items on which ninety percent of the children agreed.

3. Tests correlating highly with intelligence and having no independent value.

Thus ten tests of two forms each were left. Interesting correlations and intercorrelations were computed and conclusions drawn.

While the religious educator must be concerned with a more detailed measurement of character habits than that attempted by Brandenburg, it is hardly probable that the highly technical attempts of Hartshorne and May will become common practice for some time to come.

In the one illustration we have an attempt to judge the presence or absence of abstract personality traits. In the other, the attempt to discover specific mental skills. This is not, however, a case of the testing method versus the rating method, but rather the attempt to measure specific habits and skills versus the attempt to measure abstract personality traits, the existence of which is doubtful.

When then ought we to measure in Religious Education?

1. Mental content and skills.
2. Desires, attitudes, motives, etc.
3. Social behavior.
4. Self-control, the relation of all these factors to one another and to social self-organization.

Chapter II

Significant Developments in the Field of Measurement.

A. HISTORICAL SKETCH OF THE MEASUREMENT MOVEMENT.

Chave (1) described an early attempt at measurement during the sixteenth century which has come to us through the history of English law. In 1864 an English school-master invented a scale book for measuring school subjects.

In 1904 the publishing of Thorndike's book, "Mental and Social measurements", brought new techniques of statistics and measurement.

In 1905 the first standardized intelligence tests, (Binet-Simon) came with the creation of schools for subnormal children in France.

In 1910 Elliott developed an elaborate scheme of some hundred traits which was entirely subjective. Rugg (2) and several others worked out several ratings and correlations, sometimes with several teachers rated simultaneously, but no correlation exceeded 0.2.

Boyce followed this with a scale of forty-five qualities with ten divisions which was a bit less cumbersome, but it was purely subjective with nothing external against which to check.

(1) Chave, E.J.-"Studies in Religious Education" pp 130-132 Editors--Lotz and Crawford Cokesbury Press 1931.

(2) Rugg, H.O.-" Is the Rating of Human Character Practicable' P. 426, Journal Education Psychology Nov. 1932

Shortly after this, a seminar at Carnegie Institute of Technology under the leadership of Professor Walter Dill Scott, developed the idea of man-to-man comparison as a measure of greater objectivity, and from this came the man-to-man rating scale known as the ranking method.

In 1917 and 1918, Rugg and several others developed the man-to-man rating scale into the Army Rating Scale, a scale of five qualities, which was applied to educational work in the Detroit scale (1) for rating teachers and in the Pressey (1) card for rating pupils.

Dealing with the significant advance in measurement in 1921, Haggerty (1) refers to the extension of intelligence examinations and the advance in the use of these tests as the greatest development in the measurement movement. This very advance made necessary and called into use achievement tests and other measures of school progress. By 1921 intelligence tests had been developed for the whole school range. However, this merely pointed to heretofore unrecognized problems in the very limitations of intelligence measurement. Other factors such as industry, loyalty, honesty, tact, sympathy and cheerfulness play an equally prominent part in human life. Intelligance by itself is not enough.

(1) Haggerty, M.E.-"Recent Developments in Measuring Human Capacities". Journal Education Research April 1921 pp. 241-253

At this same time, 1919, as if in answer to the above need, came the beginnings of objective measurement of non-intellectual traits in the shape of Dr. June Downey's (1) "A Tentative Scale for the Measurement of the Volitional Pattern".

Since 1921 the emphasis has been on the measurement of character and personality traits. According to Symonds (2), the development followed roughly eight types of effort. These eight classes are:

A. Habit scales of which the Upton-Chassell (3) "Scale for Measuring the Importance of Good Citizenship" is a type. This is a scale of twenty-five qualities described in concrete terms of specific habits. A quantitative evaluation of each item is provided in the total score of one thousand points. Several judges assigned values to the descriptive statements on the basis of relative importance. For example:

The Good Citizen:

Stands for Fair Play

9. Stands for fairness in games or arguments.

(1) Downey, June-University of Wyoming Bulletin #16 pp 1-40

(2) Symonds, Percival M.-"The Present Status of Character Measurement". Journal of Educational Psychology Vol. 15 pp. 484-498

(3) Upton, S.M. and Chassell, C.F.-"A Scale for Measuring Habits of Good Citizenship"-Teachers College Record 1919

9. Protests against anyone's taking advantage of the weak, stammerers, cripples, or other unfortunate persons.

8. Defends absent persons who are unjustly attacked.

7. Does not let another pupil make wrong use of his work, such as copying from his examination or home work papers.

7. Claims no more than a fair share of his attention, expecially in the recitation period.

6. Does not expect special favors or privileges.

The observer checks the habits which are characteristic of the subject and computes the score.

B. Character or Personality Scales out of which grew the man-to-man comparison idea and the graphic rating scale. Usually this is a long list of qualities or attributes which the rator uses in judging an individual.

C. Self-assurance or overstatement tests used by Voelker (1). A blank of questions such as the following is prepared.

1. Do you know how to write any number up to ten million?

2. Do you know the name of the capital of each state in the Union?

(1) Voelker, Paul F.-"Function of Ideals in Social Education"-Teachers College Contribution to Education" #112

3. Do you know the names of the presidents of the United States from Washington to Wilson?

The individual who answers "no" marks zero after the question, the one who answers "yes" gives himself a grade of ten. The score for the whole list is computed and a prize awarded for the highest score.

A quiz test such as the following is then given and the replies are compared with the statements given in the above:

1. Give the capital of Utah.

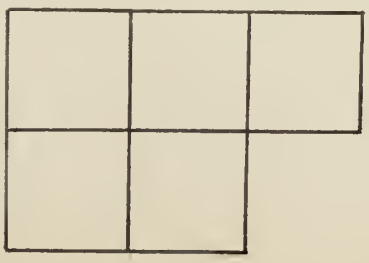
2. Write six hundred seventy thousand forty-five, and eight thousandths:

3. Name the third President of the United States.

A score of 1 is given for each correct answer as well as for each failure in the quiz test providing the answer to the first section was no.

D. Square and Circle Test, also used by Voelker.

1. Fifteen toothpicks are placed on the table so as to form five equal squares, such as:



Remove three toothpicks so as to leave three squares only. A copy of this diagram is given and the subject is instructed to accept no help but to solve it himself.

The second examiner opens a box of Gilbert puzzles and says:

"I have some puzzles here for you. Did the previous examiner give you any puzzles? Could you do them? What were they? These are easy. Shall I show you?" The subject's replies are carefully recorded.

The purpose of this test is to discover whether or not the subject can be trusted to refuse help in the solution of a puzzle which he has been instructed to try to solve independently.

The subjects who accept help are scored zero. Those who refuse help are scored ten. Those who partially yield are scored on the amount of yielding.

II. The circle test. Each subject is handed a piece of cardboard 12" by 12" on which are located five circles 5/8" in diameter arranged as an imaginary equilateral pentagon.

The subject is instructed to place a pencil mark in each circle, with his eyes closed, as his hand sweeps around the circumference of the pentagon. If he places a mark in each circle the first time around, he marks on a card, "first trial correct." If he fails, he writes "first trial wrong." Five trials are given and the individual reporting the highest number of successes is rewarded.

The feat is practically impossible of accomplishment. Each report of one or more successes is scored zero and each report of all failures is scored ten.

E. Paraffin Completion Test, also used by Voelker. This test is prepared on a four page folder. Page one contains the completion exercise, page four contains the complete exercise, page three is coated with paraffin. For example:

Page 1.

1. Boys and _____ soon become_____ and women.

2. The _____ are often more contented _____the rich.

Page 4.

1. Boys and girls soon become men and women.

2. The poor are often more contented than the rich.

The student is instructed to fill one word in each blank space. Page four is folded back and not seen. After the test is completed, the paper is folded so that page 1 and page 4 can be both seen at the same time and each student is asked to score his own paper. The teacher leaves the room and opportunity is given to cheat. Words added during the correcting process do not appear on the paraffin which holds the record of the first attempt to fill in the vacant spaces. Score zero if subject cheated, ten if he did not cheat.

F. Speed of Decision Test, used by Downey (1). This is a list of thirty opposites such as :

1. careful careless
2. daring cautious
3. ambitious unambitious
4. selfish unselfish
5. punctual tardy

Subject is asked to draw a line under the word in each pair which more nearly describes himself. Pairs taken in order. None to be skipped. Time limit 45 seconds.

G. Questionnaire, which is the oldest and most direct method, and has a high degree of subjectivity.

(1) Downey, June - "Downey Will-Temperament Test" - World Book Co., 1923 - Yonkers-on-the-Hudson.

This is a long list of questions. Inferences are made from the answers. Value questioned.

H. Tests of Judgments of moral traits, or ethical discrimination tests such as Kohs's (1) "Ethical Discrimination Test." This is a test designed to measure significant ethical knowledge and ability to make ethical and moral judgments. Six types of response are required of the subject.

a. Social Relations

What would you do when a playmate hits you without meaning to? If you found that a man had just hung himself?

b. Moral Judgment

Which is worse? Fighting, killing, hurting, quarreling, hating.

c. Meaning of Proverbs

Don't count your chickens before they are hatched means_____

d. Definitions of Moral Terms

Good means dirty, right, break or bad.

e. Evaluation of Offenses

(1) Kohs, C. S. - "Ethical Discrimination Test" - C. H. Stoelting and Co.-424 N. Homan Ave., Chicago.

For such offenses as bigamy one should be praised, ignored, scolded, put in jail, put in prison, killed.

f. Moral Problems

You should not throw hot water on a cat because; you only waste water, hot water hurts the cat, cats bathe in cold water.

For the most part this is a multiple choice test, but there is opportunity also for essay response.

The most significant recent contributions to this developing movement have been made in the measurement of character by the Character Education Inquiry (1), in the measurement of attitude by Thurstone and Chave (2), and in the measurement of certain aspects of faith in God by Donnelly (3).

(1) Hartshone and May-"Studies in Deceit"-Macmillan 1928
Hartshone and May-"Studies in Service and Self Control Macmillan 1929
Hartshone and May-"Studies in the Organization of Character-Macmillan 1930

(2) Thurstone and Chave-"The Measurement of Attitude"-University of Chicago Monograph 1929

(3) Donnelly, H.I.-"Measuring Certain Aspects of Faith in God as Found in Boys and Girls 15-16, and 17 years of Age"-Westminister Press 1931

B. MEASUREMENT IN THE FIELD OF RELIGIOUS EDUCATION AT PRESENT.

Perhaps we should say "measuring instruments in the field at present which are of interest to religious educators." It becomes immediately evident that in the measurement of character habits or personal qualities, the efforts of so-called secular educators and religious educators overlap. Some tests have been developed by agencies which regard their function as distinctly secular, and other tests have been developed by workers in the religious field, which are useable in either field. In the following paragraphs our major consideration will be with the instruments of measurement which can be used in the religious field regardless of the authorship or of the source from which they came.

It is next to impossible to list all of the instruments of measurement which are of interest to Religious Educators, partly because of the wide variety of sources from which experimentation in this field comes, and partly because much that is presented is inferior to other similar efforts. Hence the tabulation of tests, in rather arbitrary classifications in the following pages, is not in any way complete, but is, we hope, in a measure comprehensive.

1. Biblical Knowledge Tests.

A. Whitley (1) Biblical Knowledge Tests, is published in two forms, Form A 43 and B 21, and is a non-critical multiple choice test of Biblical information. To be used with ten year olds and older.

B. Church School Examination Alpha (2) is a test of seventy-five multiple choice questions in three sections. The first section deals with Old Testament information, the second section with New Testament information and the third section with Ethical discrimination. To be used with ten year olds and older.

C. Laycock Test of Biblical Information (3) is a test designed to test Biblical facts. This also is a multiple choice test, to be used with ages twelve to sixteen.

These three tests are improvements upon Giles Sunday-Examination A (4) which was a Biblical information true-false test.

(1) M. T. Whitley, Teachers' College, Columbia University, New York.

(2) Published by W. L. Hanson-Boston University School of Religious Education, Boston, Mass.

(3) Published by S. R. Laycock-University of Alberta Bookstore, Edmonton, Alberta, Canada.

(4) Now out of print.

The multiple choice method is an improvement on the true-false plan. They attempt to measure nothing but Biblical Information, although interesting correlations have been worked out between the three divisions of the Church School Examination Alpha, from which inferences concerning the relation of Biblical information to ethical discrimination have been made. However, these inferences have little value, due to the difficulty of isolating such factors as school life, and home training which also operate in the forming of ability in ethical discrimination.

2. Tests of Religious Ideas.

A. Multiple-Choice Test of Religious Ideas (1). An effort to discover the religious ideas of people. The procedure is to check the five best answers out of a possible list of fifteen for ten questions. Can be used with all persons who can read.

B. "A Scale for Measuring Certain Aspects of Faith in God as Found in Boys and Girls, Fifteen, Sixteen and Seventeen years of Age" (2). This is a four part test. Part 1 is a multiple-choice vocabulary test. Part 2 is a general questionnaire. Part 3 is a questionnaire in which the subject is

(1) Indiana Survey of Religious Education-Doran 1918 Vol. 3-pp 430-450

(2) Donelly, M. I.- A Thesis in Education.- Westminister Press 1931--pp 95-102

asked to check all of the list of fifty statements which indicate his position in answer to the question, "How Much Do You Trust God", Part 4 is a five point rating scale on which the subject rates his feeling regarding forty-one questions of belief about God.

C. Union Test of Religious Ideas (1). The purpose of this test is to measure certain intellectual concepts. A yes or no response to a series of religious questions, and a completion exercise on the story of the Bible, content and relationships, is the procedure followed.

3. Ethical Discrimination Tests.

A. Fernald Ethical Discrimination Test No. 36035 (2). The subject ranks or arranges a series of offenses from the least to greatest in order of gravity. The purpose is to get an insight into mentality and the degree of responsibility of the subject.

B. Fernald Ethical Perception Test No. 27105 (3). This test is designed merely to give some insight of a person's intelligence in the matter of

(1) Published by Dept. of Religious Education-Union Theological Seminary, 3041 Broadway, New York.

(2) Published by C.H. Stoelting Co.-424 N.Homan Ave.-Chicago

(3) Published by C.H. Stoelting Co.-424 N.Homan Ave.-

right and wrong, realizing that knowledge is little safeguard against wrong doing. Ten questions, seven to be answered yes or no, the last three by solving problems.

C. Koh's Ethical Discrimination Test (1). The test is designed to measure ethical knowledge, opinions, judgments and interpretations. It is a mixture of multiple choice, problem solution, and interpretation of proverbs, methods.

D. Union Test of Ethical Judgment (2). This test is designed to measure ethical standards with regard to current problems of life relationships such as economic justice, gambling, questionable recreation, etc. The method is a yes or no response to questionnaire and three point rating of excellent,--fair,--poor type.

4. Attitude Tests.

A. Test of Racial Attitudes (3). The purpose is to measure attitude toward people of other races. Subject chooses one of the following

(1) Published by C.H. Stoelting & Co.-424 N. Homan Ave- Chicago

(2) Published by Dept. of Religious Education-Union Theological Seminary, 3041 Broadway, New York

(3) Published by Goodwin B. Watson-Teachers College-Columbia

words: All, Most, Many, Few, No, to complete a problem statement such as: "Jews will cheat you if they can". Thirty-six elements ranging from conservative to radical in idea.

B. Hart Test of Social Attitudes and Interests (1). The purpose is to reveal dominant likes and dislikes by marking several activities plus or minus for like or dislike.

C. Thurstone and Chave's battery of attitude scales (2) four scales for measuring "Attitude toward God", "Attitude Toward Prohibition", "Attitude Toward War", and Attitude Toward the Church". Given a list of statements the subject is asked to check the statements with which he agrees, double check the statements about which he feels most strongly and place a cross before the statements with which he disagrees.

5. Character Tests.

A. *Battery of Tests developed by The Character Education Inquiry (3).

*(All tests in this battery here listed are published by the Association Press-New York)

(1) Published by Iowa Child Welfare Research Station-University of Iowa-Iowa City-Iowa.

(2) Published by University of Chicago Press-Chicago.

(3) Hartshorne and May-"studies in the Organization of Character"-McMillan 1930-pp 366-367.

(1) The I & R Arithmetic, Word Knowledge and Spelling Test attempt to measure honesty in relation to school work.

(2) The S A Test measures the tendency to misrepresent customary behavior for social approval.

(3) Tests of classroom cooperation. Measures work done for gain of self or gain of class, work sacrificed for sake of class.

Thus far, in our consideration of measuring instruments in the field of Religious Education, we have been concerned, primarily with the test type. Biblical information tests, tests of religious ideas, tests of ethical discrimination, and attitude tests and scales are primarily concerned with the measurement of knowledge. Character tests are concerned primarily with measuring the responses to controlled labratory situations.

The relationship between knowing and doing is not measured in the above instruments and as Witty and Lehman (1) pointed out, "serious temptations are seldom placed before thesubject in laboratory situations", P.412. Another question we immediately ask is whether or not we can regard the response to a laboratory situation as an indication of what the response would be in a kindred life situation.

(1) Psychological Review 1:27

These and many similar problems inevitably raise the question of policy or procedure. Should we be primarily concerned with evaluating responses in a few controlled situations, and should we be primarily concerned with knowledge, opinions, religious ideas? Should we not rather be concerned with observing and evaluating many responses to many and varied life situations if we wish to discover conduct and character habits of the subject?

This second suggested approach is properly identified with rating scales and the process of rating, consideration of which will be given in the following pages.

Chapter III

The Rating Scale.

A. THE THEORY OF THE RATING SCALE.

Thus far we have concerned ourselves primarily with the need for measurement, what ought to be measured, and a classification of phenomena which is now, in a small way, being measured. Methods of measurement, and basic differences, in instruments need now to be considered.

Width, heighth, thickness, depth, and volume can be measured objectively, by means of an objective instrument such as a rule or tape measure. Thus we say a man is six feet two inches in height and that another man is four inches shorter. The measurement is purely objective, due to the universal acceptance of the inch as the unit of measurement.

The handwriting of a pupil is measured by comparing it to a universally accepted sample of perfect handwriting. The measurement is still objective, but when we wish to know how much poorer it is than the sample, the element of subjectivity enters with the necessity of personal judgment.

If we turn our attention now to the processes of learning through which the pupil went in order to be able to write, and try to discover the degree of patience, interest, and initiative which was present in the learning process we find that this data does

not lend itself to objective treatment, and if it is to be measured at all, must be measured by the highly subjective process of judging.

Thus we have the distinction between objective and subjective types of measurement. There is no question as to the desirability of objective measurement as over against subjective measurement, but most psychological phenomena can only be measured by the so-called subjective method. The first stride toward improvement, however, is standardizing the methods of judging.

The rating scale is the instrument or medium through which we do our judging, and is built on certain basic assumptions or theories.

1. The first basic assumption is that character habits and personal qualities or characteristics exist and can be measured. Every attempt at rating is contingent upon this basic assumption. Human beings have been using comparisons ever since they have been human beings. "More beautiful", "less strong", "less happy", "more diligent", etc., are phrases which are so much a part of human vocabulary and life that the existence of qualities can hardly be denied, and yet we cannot say that they exist in the abstract.

Dealing with the fact that failure to get results

in character education is based on the theory of transfer which modern educational psychology denies, Hartshorne and May (1) point out; "In our studies in conduct we have accumulated a large amount of data which will enable us to test scientifically the truth or falsity of this theory. If, for example, honesty is a unified character trait, and if all children either hav it or do not hav it, then we would expect to find children who are honest in one situation, to be honest in all other situations, and, vice versa to find dishonest children to be deceptive in all situations. What we actually observe is that the honesty or dishonesty of a child in one situation is related to his honesty or dishonesty in another situation mainly to the degree that the situations have factors in common. - - - - - Thus we see very little evidence of unified character traits. - - - - Honesty is simply a name to describe conduct as observed in specific situations. The doctrine of specificity as shown by the facts cited above, maintains that a child's

(1) Hartshorne and May "Summary of the Work of the Character Education Inquiry" Religious Education, Vol. 5 P. 754

conduct in any situation is determined more by the circumstances that attend the situation than by any mysterious entity residing within the child." Thus the child develops many specific habits of being honest, and our study of character becomes the process of observing and judging the presence or absence and the degree of presence or absence of certain well defined types of behavior.

Thus if we wish to discover how trustworthy a child is we proceed by one of two methods. We either compute the scores on tests such as Voelker's self assurance or overstatement test, square and circle test, paraffin completion test (described on preceding pages) or we build a scale of many specific responses to life situations similar to the following and indicate a position on the line above which in our judgment corresponds to the degree of trustworthiness which the child indicates.

1. Does he take advantage of teacher's absence from the room to cause disturbance?

Always acts same as though teacher were in the room.	Enjoys fun but seldom participates.	Joins in disturbance on some occasions	Joins heartily in disturbances started by some one else.	Never misses chance to take advantage of the teacher's absences to start a disturbance.

2. Does he do his own work or does he depend on others?

Always does own work, never copies.	Does own work, but copies rather than fail.	Does own work but copies for the sake of better marks.	Does own work on occasion but would rather copy than work.	Always depends on others. Copies whenever he gets a chance.

In the one method we create specific controlled situations in which the responses are scored. In the other we indicate our judgment of the degree of the presence or absence of desirable character habits from the observation of many life situations. The one is evaluation of response to created or controlled stimuli. The other is evaluation of many typical reactions to natural, uncontrolled stimuli, or life situations.

However, it is evident that many of the basic assumptions underlying rating are also applicable to testing.

3. These qualities or habits can be so defined as to provide a working basis for discovery of their presence or absence in many diver situations and the degree in which they are present or absent. A problem presents itself. It is usually stated in the form of criticism in two ways.

First, in the question of the overlapping of qualities. Secondly, according to Chave (1) "---------the terms used to describe social and religious attitudes, tendencies, and qualities of activities, and accomplishments are not uniformly interpreted." Life is not lived in compartments, but as a unit. Hence when we say a boy is truthful and honest in a given situation we immediately face the question, where does honesty leave off and where does truthfulness commence? An overlapping is evident, but is the overlapping in qualities or in titles we have attached to one quality? If a boy is truthful he is also honest. What is the difference if we are measuring the same quality and calling it honesty in one situation and truthfulness in another? Are we not more anxious to discover the presence or absence of the quality than the title? Suppose we wish to discover whether or not a boy does his own work in school. We build a questionnaire and have him check himself or we observe his conduct at school. We discover that in some situations he relies on himself, while in others he copies.

(1) Chave, E. J. -"Supervision of Religious Education" P. 313. University of Chicago Press 1931.

We say he is not perfectly honest, nor is he perfectly dishonest and assign him a position on a scale or give a standing in comparison to others, or place a numerical value after his name. This stands for the degree of the quality that we prefer to call honesty. At any rate we have measured the quality or habit. By defining the type of quality we wish to measure in carefully selected descriptive terms we are not necessarily defining honesty, but the habit or quality we wish to observe and placing it in the category of honesty. If we try to measure honesty in a general way we are immediately conscious of "overlapping" or lack of uniformity in interpretation of the term, but if we are careful in our descriptive definitions of what we wish to measure, the problem diminishes and we have at least a working basis on which to proceed with our measurement.

3. The third basic assumption is that some qualities or habits are present in greater degree in the activities of some individuals than they are in others. If this were not so there would be no need for rating. There would be no persons at the extreme ends of the scale, and therefore there would be no need of a scale. Two clerks work side by side in an office. One comes to work with clothes pressed,

shoes shined, hair combed, clean hands and nails while another comes with clothes unpressed, hair untidy, shoes dusty, hands and nails unkempt. One clerk places his papers neatly away at closing time, the other tosses them hurriedly into a drawer. We say immediately that one is neater than the other. We cannot say, however, that one is neat and the other is not, because there are easily recognizable varying degrees of neatness. The same observation might be made of ambition, friendliness, intelligence, sympathy, cooperativeness and similar qualities or characteristics. This leads us directly to our fourth basic assumption, which is implied in the above illustration.

4. The presence or absence of these qualities and the degree of their presence or absence may be inferred from behavior. As has already been pointed out, character is the composite of many specific habits of being honest, truthful, etc. It has also been shown that the relationship between specific habits of being honest, etc., is determined by factors operative in various situations which are similar or common. Thus it becomes evident that the observation of the reactions to a few situations is not an adequate index of a child's honesty.

A boy may be trustworthy in running errands and handling money for his mother. He may never cheat in school. He may tell the grocer he was given too much change. At the same time he may spend part of the money given him as

offering for Sunday School or he may ride on an out-of-date transfer on the street car if he can get by with it.

Observation of a few of this boy's activities and the rating of honesty or trustworthiness will probably not reveal habits and practices which ought to be observed in order that the most accurate picture of character might be obtained.

Thus we cannot generalize from too few phases of behavior. Many types of activity and the reactions to many different situations must be carefully observed if our ratings are to be valid estimates of that which we attempt to measure.

5. Fifthly, it is possible for observers to distinguish between the degrees of the presence or absence of a given quality or habit. Five men stand in a row applying for a position. 'A' is punctilious in his dress, his hair is oiled to stay in place, a flower adorns his button hole, his clothes are pressed with knife-like severity, hands daintily manicured--almost foppish in appearance. 'B' is dressed simply, hair combed, clothes pressed, clean hands, shoes brushed but not shined. 'C's' clothes are rather worn, carefully brushed, hair combed, cleanly shaven, hands clean, nails unkempt. 'D' has clothes of good quality,

unpressed, hair not combed, shoes in need of a shine, hands and nails unkempt. 'E' has soiled shirt, tie knotted carelessly, unclean, clothing soiled and unpressed, shoes badly in need of a shine. With A, B, C, D, and E, standing together, it is easy to infer which of the five is neatest and least neat. It is also easy to place B as next to A and D next to E with C in the middle. It is not always as easy as this. Very often the qualities are not as easily discernible and often there is real question as to which of two or more individuals should be placed at one end of the scale and which of several at another, with still greater question about those in between.

6. The sixth basic assumption is that the assigning of numerical values to the degree of the quality which, in the judgment of the observer, is apparent or the assigning of the degree of the quality to an area on a linear continuum, makes the judgment more concrete. Rugg (1) in this application of the principle of man-to-man comparison to a rating scale for high school students, assigned the following numerical values to each of the five groups or ranks of students; the best student -3a, better than over one

(1) Rugg, H.O. - "Rating Scale for High School Students" P. 451 Journal Educational Psychology Vol. 12: 1921

35, average-22, poorer than average-14, poorest student-6. In this way the total judgment of the pupil is facilitated. Student 'A' might rank 6 in the first quality, 35 in the second, 20 in the third and so on. The final score then shows more concretely the presence or absence of various qualities and the degree of the presence or absence of the desirable and undesirable qualities due to the universal recognition and interpretation of numerical values. However as Chave (1) points out, the use of numerical values must be used with caution. "For the sake of convenience numerical indexes are used to describe differences, but the mathematical language does not mean that exact measures have been taken. Measures are all subjective estimates and rough approximations, findings being interpreted by reference to the scale used. Though scores on a knowledge test may be given in precise numbers, such as 94 or 96, or a person's position on an attitude scale as 3.7, it does not mean that we have an exact measure of his knowledge or attitude as the mathematical symbols seem to suggest. All we have are symbols of relative position, and the meaning of

(1)"Studies in Religious Education" Cokesbury Press 1931, Page 130

then depends upon a thorough understanding of the way in which the figures have been obtained. This is, however, the same situation as in general education when we state a person's I.Q. as 140. It means the person has a superior intelligence as far as the test given has been able to measure differences, but the inference often is that when a person has been able to answer a number of questions in an approved way that he has thereby revealed his native mental capacity. The I.Q. may be a convenient index, but it is by no means an exact measure of intelligence."

7. Seventhly, by carefully defining and thus restricting by definition the types of behavior to be observed, the judgments of the presence or absence of a quality and the degree of presence or absence, will have greater validity. A scale is constructed to measure the spirit of cooperativeness of kindergarten children. After several independent ratings by several competent judges, correlations between judgments of different teachers and different judgments of the same teacher are taken and they show a coefficient of correlation which is theoretically perfect. Have they measured the cooperative spirit of the child and will they say this child is very cooperative? The judgment of the cooperative spirit of the child is

valid for the school room and other situations in which there are factors in common with the school room situation and not for the general spirit of cooperativeness.

(The above seven items are the basic theories or assumptions underlying rating as a method of measurement. This last theory has to deal with <u>how</u> we are concerned with rating as measurement.)

8. In the last place, some qualities or habits on the basis of race experience, social opinion, and religious belief, are considered highly desirable and can be cultivated, while others are undesirable and can be sublimated. Jealousy, envy, selfishness, untrustworthiness, and similar characteristics appear oftentimes and are recognized to be distinctly anti-social, while, on the other hand, punctuality, honesty, altruism, and similar qualities are regarded as highly desirable. The concept of education as a process of producing changes is without value unless we recognize desirable qualities or ends and thus give direction to the process. To what extent are desirable qualities present at a given time in the process? Is the procedure which has been selected, achieving the desired objectives? In what way can we discover answers to the above and similar questions? We must have some way

of collecting data to serve as the basis of judgment and we turn to the rating scale as the instrument which is best suited to this purpose. However, the question arises; Can we not discover the answers to the above and similar questions by the use of tests?

The most careful and comprehensive testing program to date has been conducted by the Character Education Enquiry which tabulates its findings in three volumes (1). Almost eighty tests were used as the basis for the measurement of knowledge, attitude, intelligence, conduct, background, and social adjustment. On the basis of this very extendive study the authors came to the following conclusions (2).

1. "Character cannot be measured adequately by any single or simple test that can be administered in one hour and scored in ten minutes.

2. No algebraic summation or average of any set of test scores, no matter how extended or elaborate, will give a true index to character.

(1) Hartshorne and May - Studies in Deceit - McMillan 1928
" " " - Studies in Service and Self-Control - McMillan 1929
" " " - Studies in the Organization of Character - McMillan 1930

(2) Hartshorne and May - "Summary of the Work of the Character Education Inquiry" - Religious Education Vol. 5, P. 61

3. If a large number of samples of conduct, knowledge, attitude, intelligence, background, and social adjustment are taken, and if the general algebraic level for each individual is determined and at the same time if the variability of each individual's scores around his own mean is computed, a combination of these two values will indeed yield an index or score of character."

These conclusions point to the fact that simple tests or few tests do not provide an index to character, and that the only valid index to character is secured by the combination of many and varied types of tests. Furthermore, this experiment required approximately thirty hours of time from each student tested.

Thus, we see that a valid testing program is impractical because: in the first place, it limits the use of this method to a very few trained experts in the field , and in the second place, the amount of time required makes it impossible of use in the field of religious education, where, for the present at least, the time spent is only approximately fifty-two hours per year.

Hartshorne and May (1) further point out that "An interesting practical implication of all this for character testing is that since it is probably easier

(1) Hartshorne and May - Religious Education Vol. 5 -- Page 617

to secure a series of valid judgments concerning attitudes and conduct tendencies than it is to secure an equally valid series of objective tests, and since the theoretical correlation between the two is almost perfect, teachers and others interested in securing a character score on children might very well look forward to doing so by securing a large number of reliable observations and ratings.

In the second place, assuming that the attempts of the Character Education Inquiry have completely measured character as far as honesty, service, and self-control are concerned, there are still a large number of intangible qualities such as tolerance, cooperativeness, resourcefulness, courage, etc., for which there are as yet no valid, objective tests.

Thus, because of problems arising in the use of tests, and because of the fact that there are not as yet valid tests with which to measure all of the qualities which should be measured, the rating scale, which is an instrument to be used in the observation of an individual in many situations and to record the mental summations of the rator of many subjective judgments of behavior of the individual observed, it is the most usable instrument available.

B. TYPES OF RATING SCALES.

The basis for a consideration of types of rating

scales must be more or less arbitrarily chosen. Barr and Burton (1) list two types of rating scales, one brief or compact rating scale and one elaborate scale. This distinction is evidently made on the basis of useability, and is hardly the basis of a discussion of types of rating scales.

Nor is the difference of methods in rating a satisfactory basis for a discussion of types of scales. Freyd (2) lists eleven distinct methods of ratings. Oftentimes two or more of these methods may appear in one scale.

Symonds (3) assumes two types of scales in his discussion of "Rating vs Ranking", i.e., the rating scale as a type and the ranking scale, or man-to-man comparison as a type.

The basic purpose of the rating scale, and the underlying theory is much the same in all scales, but in the evolution of rating, motivated by a desire for greater validity, peculiar types have emerged. Differences in methods of scoring, classification of qualities or habits and similar bases of analysis

(1) Barr & Burton - "The Supervision of Instruction" pp. 468-480. D. Appleton & Co., 1926

(2) Freyd, Max - "The Graphic Rating Scale" pp. 83-102 Journal of Educational Psychology. Vol. 14, 1923.

(3) Symonds, Percival M. - "Notes on Rating" Journal Applied Psychology. Vol. 9, pp. 186-195

will hardly determine type, but differences in principles of construction are the basis for the classification of types of scales.

There appear to be four types of scales, i.e., the simple rating scale, the graphic rating scale, the man-to-man comparison scale, and the score card.

1. The simple rating scale.

The simple rating scales now in use are adaptations of the Boyce (1) scale or efficiency record. Scales of this kind are used primarily in the supervision and administration of teachers and in self-rating of subjects. A list of desired qualities or habits is selected. Each quality is described in detail by a list of sub-heads, each of which is descriptive of some factor or phase of the quality to be measured. Thus by breaking up the quality into a series of parts, the judgment is not a lump judgment but a series of concrete judgments which are more accurate than a general estimate of a quality would be. Sometime these sub-headings are stated as questions, which aids in making the judgment more concrete. Descriptive concrete questions also add validity to the test by accurately defining, in universal terms, the phase of the habit which is to

(1) Boyce, Arthur Clifton-"Methods of Measuring Teachers' Efficiency"--14th Yearbook Nat'l. Society for the Study of Education part 2, PP 44-45.

be observed.

Three or five columns are then prepared and the rator is asked to place a check mark in the column which, in his judgment, indicates the degree of the presence or absence of the trait being observed. The columns may be headed good, medium, poor, or A, B, C, may be used. In a five point scale one may use very good, good, medium, poor, very poor, or A, B, C, D, E, or 1, 2, 3, 4, 5, etc. Numbers are often applied to each of the five columns and a numerical score is computed.

The following sample is from a scale which is used in the University of Chicago Plan (1) of supervising and administering practice teaching, and is a good example of the simple rating scale.

1. Skill in Teaching technique
   Preparing instruction materials
   Attending to individual needs
   Developing independence of pupils
   Directing pupil study

2. Classroom management and school routine.
   Attention to routine matters
   Management of pupils

The supervising teacher places a check mark in the column which, in her judgment, represents

(1) Breslich, Gray, Pieper and Reavis-"The Supervision and Administration of Practice Teaching" Ed. Adm. & Sup. Jan. 1925 pp 1-12

the degree of proficiency the student has reached. In this scale F means failure and A means perfect.

Brown (1) uses the simple rating scale plan in a self-rating scale for students, a sample of which follows:

Punctuality

		Ex.	VG.	G.	F.	P.
1.	Do I keep my appointment?					
2.	Do I obey my parents and teachers promptly?					
3.	Do I perform unpleasant tasks promptly?					
4.	Do I return borrowed articles promptly?					
5.	Am I economical of my time both at work and play?					
6.	Do I arrive at school on time?					
7.	Do I hand in written work on time?					
8.	Do I go to bed and get up at regular hours?					
9.	Do I get to meals on time?					
10.	Am I prompt and gracious in acknowledging kindness?					

The pupil places a check mark in the column which most nearly fits him.

Furst (2) suggests a method of scoring on a five point scale. He selected six traits; character, mind, force, social qualities, knowledge, and technique

(1) Brown, Edwin J-"A Character Conduct Rating Scale for Students"- Education, Vol. 50:369-379.

(2) Furst, Clyde-"A Simple Literal Personal Rating Scale" P 463 Educational Administration and Supervision Nov., 1922.

each with twelve to eighteen descriptive sub-heads, and used the vowels a,e,i,o,u, to represent degree from best to worst. By combining the first letter of the word denoting the trait with the vowel representing the column in which the quality was checked, the score of the person rated is read in literal fashion for the perfect individual, CA, NA, FA, SA, KA, and TA.

There is no particular value in any method of scoring. One method is as good as another if the judges and the subject understand perfectly the meaning of the scores.

Rugg uses this type of scale in his "Rating Scale for Pupils Dynamic Qualities (1), form A. The headings; Ability to learn to assimilate new ideas, qualities of Industry and Attitude toward school work, Qualities of Leadership, Teamwork qualities, Personal and Social qualities; with from 6 to 10 appropriate sub-headings were arranged on a five point self-rating scale.

A three point scale of five characteristics is in use in the public school system of Duluth (2), while a five point scale of five characteristics is in use in

(1) School Review Vol. 28:337-349.

(2) Bracken, John L.-"The Duluth System for Rating Teachers" Elementary School Journal Vol. 23: 110-119.

the schools of Madison, Wisc. (1) as supervisory instruments. Both are excellent examples of the simple rating scale.

2. The Graphic Rating Scale.

While the purpose of the graphic rating scale like that of the simple rating scale, is to judge, from the behavior of a subject, the presence or absence and the degree of presence or absence of desirable qualities, the construction of the graphic scale, as an instrument, is very different from that of the simple scale. A straight line is drawn which represents the range of the quality or trait which is being observed. Using 'honesty' for the sake of illustration, one end of the line is to represent extreme dishonesty while the other end represents extreme honesty. If the scale is to be a three point scale, at each end of the scale under the line are printed descriptive phrases indicative of the two extremes of the quality. Beneath the middle of the line are printed descriptive phrases indicative of the neutral or mid-position of the trait. The rator merely makes a check mark at the position on the line which represents to him the position of the subject with regard to the trait observed.

(1) Giles, J.T.-"A Recitation Score Card and Standards" Elementary School Journal Vol. 23:25-36.

The following is a sample of the graphic rating method taken from a graphic rating scale of seventeen qualities by Freyd (1).

1. How does he impress people by his physique and bearing?

Excites admiration. Very impressive.	Noticeable for good physique and bearing	Makes satisfactory impression.	Unimpressive physical bearing.	Arouses repulsion. Looked down on.

The rator places a check on the line at the point which, in his judgment, indicates the position of the one rated with regard to physical bearing.

Freyd (2) lists thirteen rules for constructing a graphic rating scale.

1. Decide on the extremes of the trait--one extreme of a scale may have several opposites..

2. The line should be of such length that a stencil for scoring the ratings can be easily calibrated.

3. No breaks or divisions in the line.

4. The line should not be more than five inches in length, so that it may be grasped as a unit.

(1) Freyd, Max-"The Graphic Rating Scale" pp. 83-102 Journal Educational Psychology Vol. 14; 1923.

(2) Freyd, Max-"A Graphic Rating Scale for Teachers" Journal Ed. Research Vol. 8 pp. 433-39.

5. Not more than five descriptive phrases nor less than three.

6. The end phrases not so extremely worded as to be seldom used.

7. Phrase descriptive of the average degree of the trait should be in the middle of the scale.

8. If there are five items, the intermediate ones should be closer in meaning to the center one than to the extremes.

9. Only universally understood phrases should be used.

10. Use in place of "average, fair, excellent, etc.," adjectives, which, in themselves, express varying degrees of weight, (for "extremely neat", use "fastidious", etc.)

11. Descriptive phrases short and to the point.

12. Phrases set in small type with plenty of white space in between.

13. Favorable extremes of the scale should be alternated so as to eliminate the motor tendency to check on one side of the page."

Certain virtues of the graphic method immediately become apparent. It is simple, concrete, and useable. The recognition of universal descriptive phrases and the fact that no quantitive measure is asked for,

make it easily handled and quickly scored.

Freyd (1) has used this principle in the construction of a scale for measuring qualities which make for success in teaching. This scale attempts to measure seventeen qualities. He reports considerable interest on the part of teachers.

3. THE MAN-TO-MAN OR RANKING SCALE.

The procedure in the man-to-man method or rank method is somewhat different than either of the above two types. The simple scale and the graphic scale are constructed to enable a rator to judge the presence or absence and the degree of presence or absence of qualities in an individual. The man-to-man method is essentially the comparison of individuals on the basis of the traits or qualities we wish to measure. The construction of a man-to-man scale would probably be done in this fashion ( ).

a. A list of names of the group to be measured would be compiled.

b. The desired qualities, completely and carefully defined would also be listed.

(1) Freyd, Max - "A Graphic Rating Scale for Teachers". pp. 433-439. Journal of Educational Research. Vol. 8:1923.

(2) Rugg, H.O. - "Self-Improvement of Teachers through Self-Rating". Elementary School Journal. Vol. 20; pp. 670-684.
"Is the Rating of Human Character Practical" Journal Educational Psychology, Dec., 1921; 425-451

c. Subjects in the list would then be arranged from best to poorest for each quality or trait.

d. Arrange persons as accurately as possible in five groups.

e. Select one person from each group as the best in each group to occupy the position on the scale.

f. Assign to the scale people the following values, ranging from best to poorest; 66,30,22, 14, 6.(1)

When the scale is used, each individual in the group is compared with the scale persons on the five steps of the scale and is assigned to the scale man, or scale step where he rightfully belongs. By assigning a quantitive numerical value to each person, the total score for the whole test is easily computed. For instance, John Smith is the most honest boy in the group, and we place 66 after his name. Bill Jones is above average but is not as honest as John. We give him 30. Jim Brown is extremely dishonest and we give him 6, etc.

(1) Rugg, H.O. - "Self-Improvement of Teachers Through Self-rating". Elementary School Journal.Vol.20; pp. 770-804. "Is the Rating of Human Character Practicable". Journal Ed. Psy. Dec., 1921; 485-501.

We find that Andy Johnson is not as honest as John Smith, but is more honest than Bill Jones, so we give him 34, and so on, until every person has been compared with the scale people.

Scale people will probably differ for every quality. Bill Jones may be the scale man for the next lower step than the top in honesty, and may not be a scale man for diligence, patience, and may be the scale man at the lowest point for tact.

Miss Chassell (1) used the man-to-man ranking method with kindergarten children. A sample of the scale.

I. Habit of Work.		Names of Scale People.
a. Ability to initiate projects and carry them out.	25	
b. Ability to fail and persevere.	15	
c. Ability to carry out directions of others.	5	
II. Participation		Names of Scale People
a. Ability to contribute to the social development of the room.	5	
b. Ability to take an intelligent interest in the social activities.	15	
c. Ability to participate and be responsible for the social organization.	25	

(1) Chassell, Clara F. - "The Army Rating Scale method in the Kindergarten" -- Jour. Ed. Psy. Vol. 15; pp. 43-52

Barr and Burton (1) point out that some critics of this method "feel that it would be difficult to construct a scale using the names of teachers so that very many people could be familiar enough with the names used, to apply the complete card." It is also suggested that criticism from teachers condemns this method on the ground that it rates the doer rather than the deed.

Rugg (2) says of this method, "The task of comparing one person's qualities with anothers is fraught with so much difficulty as to be impractical in rating the rank and file of persons and for most practical activities of life." And yet he uses this method in his, "Rating Scale for Pupil's Dynamic Qualities" (3) and his self-rating scale for teachers (4).

Haggerty (5) describes the Scott comparison scale as the first significant and ambitious effort and

(1) Barr and Burton - "The Supervision of Instruction" Appleton 1926--P. 480.

(2) Rugg, H. O. - "Is the Rating of Human Character Practicable" Jour. Ed. Psych. Vol. 13. Jan. 1922. P. 30.

(3) School Review - Vol. 28. PP. 337-349.

(4) Rugg, H. O. - "Self-improvement of Teachers Through Self-rating". El. School Journal Vol. 20 PP. 670-684.

(5) Haggerty, M. E. - "Recent Developments in Measuring Human Capacities" Jour. Ed. Research - April 1921 - PP. 241-253.

Scott says of it, "where instructions are followed closely, the results show a high degree of accuracy and uniformity. --because the Rating Scale calls attention separately to each of the several essential qualifications for an officer, it lessens the danger that judgments may be based on minor defects with a disregard for corresponding virtues."

Thus we see criticism and defense of the rating method. I believe we are safe in saying that in some situations, the intelligent handling of this method in the evaluation of the more tangible personal habits will yield a fairly reliable score.

c. The Score Card

The score card is a form of rating scale in which the various items are weighted by preliminary judgments of a group of persons whose standing is recognized and whose experience in the particular field with which the score card deals is of special value. Points are distributed for different phases of the subject and directions are given as to the basis of the rating. One of the most interesting forms of this type of instrument is the 1000 point scale published by the Inter-Church

(1) Scott, W. D. "The Rating Scale"--Psychological Bulletin-- Vol. 15. P. 404

World Movement called "Standard for City Church Plants," a sample of which follows:

I. Site----------					130
A. Location			55		
1. Accessibility	30				
2. Environment	25				
B. Nature and condition			30		
1. Drainage and soil	15				
2. Unkeep of site.	15				
C. Size and form.	45		45		

The descriptive standard or the directions which are designed to aid the scorer in his estimation are as follows:

I. Standards Involved in the Site of a Church School Plant.

A. Location
1. Accessibility
a. Near enough to the business section of the city to profit by the convergence of roads and car lines, if a "downtown" church.
b. In the direction of the city's growth rather than behind it.
c. Located centrally with respect to its entire constituency.

2. Environment
a. Adjoining attractive, clean and well-kept property (trees, lawns, etc.)
b. Sanitary and healthful,--free from malodors.
c. Remote from fire dangers,--not adjoining to large wooden or non-fire-proof buildings, gas tanks, or other fire spreading structures.
d. Quiet,--not adjacent to any factory, planning mill, or plant employing machinery or shops such as tinsmiths, auto repair shops, passing street cars or railroad trains. Streets should not be brick or cobblestone.
e. Not near overtowering buildings, but placed in proper architectural setting on a strategic location.

B. Nature of Site and its condition
1. Drainage and nature of soil.
a. Natural slope preferred, sloping away from building at a minimum slope of 1 inch in three feet.

b. Entire site should be thoroughly tiled with special provision for the basement. Protected from surface water from higher contiguous ground. Nature of soil should determine the depth of the tile.
c. Sandy loam and fertile enough for good lawns and landscape gardening.
d. Playground, quick drying, (rapidly drained) with turf or artificial surface of crushed stone or gravel. Natural soil preferred to artificial.

2. Unkeep of Site.
a. Entire site should show evidence of proper maintainance. Lawns should be well kept; shrubbery well trimmed; walks clean and in good repair; fences or walls in good state of preservation. Grounds should be free from unsightly ash piles, waste paper, rubbish of any kind and weeds.

C. Size and form of site
a. Should be large enough and of a shape to allow for the proper placing of building or buildings and for future additions.
b. Should be large enough to provide.
(1) In front for ample lawns and shrubbery for outdoor fetes, pageants and other festivals.
(2) In rear for playgrounds, tennis courts, ball ground, and other athletic facilities to be provided.
c. A plot from 3 to 10 acres depending upon the size of the community to be served is necessary for these activities.
d. Where city congestion is such to prevent acquisition of standard site, roof garden should be planned for festivals, song services, play and other activities. Its construction should care for the following elements:
(a) Adequate roof covering, rail protection, shield against wind, rain and snow.
(b) Storage facilities and the extension of all service systems to the roof garden.
(c) Special equipment consisting of tables, chairs, portable stage and piano.

e. Where playground and athletic field are separated from the church site, they should not be so distant that the school and gymnasium equipment cannot be used.

The score [illegible] in the final figure next to the amount assigned to the factor under observation, an amount which, in the judgment, corresponds to the degree in which the building under observation conforms to the standard. Each element on the card is taken singly and is scored in the light of the descriptions in the standard. The final score is then computed for all the factors observed which stands for the degree to which the church under observation conforms to the standard.

Thus we see that the score card, as an instrument of measurement, has much in common with the graphic rating scale which we have been considering, but there are also differences between the rating scale, as we have discussed it, and the score card.

In the first place, the score card provides a convenient method of recording judgments which is perhaps as concrete as any rating scheme can be. For instance, if the value 30 is given to accessibility and the actual score of a church is fifteen, the average person looking at the scale would have a more concrete idea of how nearly the church compared with the ideal, than he would if the scale read very good, or poor, or a check mark showed on a graphic scale, due to the universal recognition of numerical values and the relationship of fifteen to thirty. This, of course, does not mean that the judgment is as accurate as the number fifteen indicates, due to the subjectivity entering into the scorer's judgment as to how

nearly the church under observation approximates the Standard, but it does mean that it is recorded in universally understood terms.

In the second place the descriptions of the standard which are used as a guide to rating, are very similar to the descriptions of divisions on a graphic rating scale. However, much more detailed description and definition is possible in the use of a score card and standard than it is in the use of another type of rating scale.

In the third place, different values are ascribed to different qualities on the scale, indicating that some elements to be scored are of more importance than some others. This can be done with thoroughly objective data such as the Standard for City Church Plants deals with, but we can hardly assume that some of the more intangible elements of character are of more importance than others, much less the degree to which they are more important. One could hardly say that diligence in a certain situation was of more value than courage in the same situation, much less how much more valuable. This leads to an interesting question as to where the dividing line is between tangible and intangible qualities.

Yepsen (1) presents a score card for personal behavior in an attempt to state objectively the social adaptibility of the individual. At the time of this writing, however, no satisfactory method of scoring had been reported.

(1) Journal Applied Psychology Feb. 1929, PP 140-147.

A sample of the card follows:

Attitude of Others Toward Him

Choose him as leader.

Accept him as leader.

Play with him occasionally, not often.

Seek his companionship.

Ignore and shun him.

Accepted readily as one of the group.

Butt of crowd pick on him.

Here again the data is more or less objective inasmuch as actual observable reactions of a specific nature are the basis of the study. At least it is possible to get a rather general agreement concerning them.

In the fourth place, agreement on the part of experts in the development objective standards is possible for the more objective data while it is impossible with the more subjective data in the field of morals or character. Hartshorne and May (1) discovered a wide difference in codes between people of one type or social strata and another.

Thus the same principle of rating is used with the score card as with other types of rating scales, except a carefully prepared objective standard is prepared against which to judge and against which to score.

C - Criticism of Rating Scales. Certain inherent weaknesses in the matter of rating invite a great deal of criticism which classifies itself under three general headings: first, the question of the

(1) "Testing the Knowledge of Right and Wrong" Mono raph 1928 PP 31-32

existence of traits is a question which reflects the influence of varied psychological theories. Secondly, criticism from the standpoint of facility in the administering and scoring of a given scale. Third, the problem of the validity of the scale and the reliability of the scores. The first is a matter of psychological definition, the second is non-statistical criticism, and the third is statistical criticism.

1. The existence of traits.

Are there such things as character traits? Are we trying to measure something which actually exists? These are questions commonly asked by behaviorists. Three studies of this question (1) have been made and certain conclusions become apparent as the result of the findings of these studies.

The psychological viewpoint of the individual facing the problem will determine the answers to the above two questions. If we hold, with the faculty psychologist, that character is native, then we must accept the existence of general character traits. If, however, we hold with the behaviorist, that character is acquired through the influence of experiences in life situations, we question the assumption on the grounds that character is not distinguishable apart from the acquiring process, or apart from

(1) Symonds, Percival M.-"The Present Status of Character Measurement", Journal Educational Psychology Vol. 15: [illegible]-[illegible].

Witty, P.A. & Lehman, H.C.-"The So-called General Character Test", Psychological Review Nov., 1927 pp. 401-413.

Brown & Shel[illegible]dine-"A Critical Study in the Objective Measurement of Character", Jour. Ed. Research Nov., 1928 pp. 290-29[illegible]

situations. This same reasoning, however, that the qualities of honesty, conscientiousness, etc., are non-existent in all situations, as it has been [illegible] non-existent [illegible] like situations. Hence a [illegible] [illegible] the validity of a general character test.

Witty and Lehman (1), classify tests of moral character in three types: (a) Those involving reactions to laboratory or class-room situations; (b) Those involving reactions to hypothetical questions regarding moral situations; (c) Those involving reaction to life situations.

Their conclusions are; that the fact that the relationship [illegible] and [illegible] is not measured in the first type of test, character is not measured; that real [illegible] exist in life [illegible] situations, hence, character is not measured, due to the multiplicity of ways in which a trait manifests itself.

The last two paragraphs in the article, [illegible] a summary of their position, read as follows: ([illegible]) "If morality then is acquired, and acquired in terms of specific habit formation, our task in the schools is that of teaching children to choose intelligently those habits which will function for the good of others.

(1) "The So-called General Character Test"
Psychological Review-Nov., 1927. pp. 401-41[illegible]

A more accurate conclusion to this article would be, that general character tests are undesirable and that the only valid type of character test is one which involves life situations. This is the conclusion Brown and Shelmadine (1) reached after classifying the available tests in four types; (a) Scales in which a person is either rated or rates himself for specific traits; (b) Tests of temperamental traits which measure mental reaction in a controlled situation; (c) Pencil and paper tests; (d) Series of controlled situations such as the Woelker tests.

Can we not say then that qualities of character do exist in life situations; and can we not formulate two principles for the construction of a character scale?

A.-"Every item in the scale should be defined in terms of action. We can judge one's possession of a given trait (or quality) only on the basis of its outward manifestation. Objective considerations should enter into all our ratings. From these we make inferences concerning subjective qualities.

B.-Personal qualities or traits are manifest only in appropriate situations. If there is no opportunity for exercise of the trait in question, the rating on

(1) "A Critical Study in the Objective Measurement of Character". Journal of Educational Research, 1929 pp. 290-296

that trait is worthless (1).

2. Non-statistical Criticisms.

Non-statistical criticism includes the selection of traits, ease of administration and scoring, simplicity of the rating scale, time required, and the comprehensiveness of the task. Often times the list of qualities to be observed is so long that an overlapping of qualities is immediately apparent, and the scale becomes complicated and formidable. Again the method of scoring by the use of a stencil, a master scale or a statistical formula limits the use of the scale to a few experts in the field of measurement. Also the selection of the qualities is often not wise, in that they are not of universally recognized importance, and the descriptive terms are not universally recognizable.

These criticisms point to the need of principles in the selection of traits and the determination of methods of scoring.

(2) 1. The list must be relatively short so that the student (or rator) shall not be lost in the maze of different qualities.

(1) Kaplan, L. Marlin-"General Principles and Results of Rating Trait Characteristics"-Journal of Educational Method-Vol. 4: pp 1074-1083

(2) Brandenburg, G. C. & Remmers, H.H.-"Rating Scales for Instructors"-Educational Administration & Supervision-Vol. 13: 399-406

(1) 2. Traits must be such as are generally agreed upon by competent critics as most important.
(2) 3. It is essential that every term in the rating scale be defined as unambiguously as possible. If the contents of a given term are too varied, comparable ratings are impossible.

4. The method of scoring must be simple and easily grasped.

3. Statistical Criticism.

Questions, which arise in the use of the rating scale, include the following: To what extent does the score of a scale measure the existence or absence of the true qualities of an individual? What causes variability in independent ratings? How can the reliability of the ratings of an individual be increased? What are the problems which arise in rating?

In some types of rating scales there is an error in the score due to the inability of the rator to clarify and formulate his reactions to a given situation. If the scale is so constructed as to enable the rator to select a phrase which most nearly describes his reactions, rather than to make a spontaneous response, this error

(1) Brandenburg, G.J. & Remmers, H.H.-"Rating Scales for Instructions"-Educational Administration & Supervision-Vol. 13: 379-406.

(2) Hughes, W.M.-"General Principles and Results of Rating Trait Characteristics"-Jour. Ed. Method: pp 1924-1925

is diminished.

Another error in judgment is caused by an inadequate understanding of what is being measured. Descriptive phrases must be fool-proof, phrased in universally accepted terms and specific. This is also largely taken care of in the construction of the scale.

There are, however, at least four personal factors in judging which cannot be eliminated in the construction of the scale. They are as follows; (a) different standards of excellence held by different rators; (b) differring abilities to distinguish degrees of existence or non-existence of the trait under observation; (c) the influence of friendship or animosity, and already formed prejudices for or against the subject on the part of the rator. This is commonly known as the influence of 'halo'; (d) the lack of knowledge of the subject on the part of the rator which is very apt to warp the score.

a. Unfortunately for rating, we are not all turned out of one mould. The varying background and experience of individuals produce different standards of judgment, different mental and cultural habits, different psychological background. Hanna (1) found a wide variability between the judgments of teachers, teaching

(1) Hanna, Joseph V.-'Variable Factors Encountered in the Rating of Students'-School Science & Mathematics Vol. 25: 481-488

different types of subjects, in the rating of Junior college Students on application, ability to organize, accuracy, punctuality, aggressiveness and social qualities. He found less variability between the judgments of teachers of the same subject background. The wide range of religious belief, the large percentage of people who scoff at religion and the environmental differences make this immediately apparent in the judgment of character qualities and moral habits. What is immoral to one person may be non-moral to another.

b. Different abilities to distinguish in degree points immediately to the need of competent judges. A competent judge is one who:

(1) Is familiar with the scale or instrument,

(2) Understands and is in sympathy with the idea of measurement by rating,

(3) Is familiar with and agrees upon the qualities under observation,

(4) Is vitally interested in the qualities measured,

(5) Has had sufficient practice in rating to have become proficient,

(6) Knows the subject rated well enough to interpret his natural reactions to situations, but not so well that he will be influenced by "halo".

c. The influence of "halo". A rator often rates a

subject high if the subject is well known and well liked by the rator, or vice versa. This does not mean that the rator is dishonest, but it does mean that the feeling of the rator for the subject influences the judgment. This is not restricted to the matter of friendship. Often a teacher will allow the scholarship or academic standing of a student to influence the judgment on a general quality.

For instance, Knight and Franzen (1) report a correlation of 0.95 between moral character and quality of voice, in a judgment of several prospective teachers. Obviously moral character has nothing to do with quality of voice nor has quality of voice anything to do with character. This means that because the ratings of moral character and the ratings of quality of voice were so nearly "alike" or "identical" quality of voice, if pleasing, influenced the judges to rate the subject high in moral character, and if displeasing, to rate the subject low in moral character. Thus quality of voice influenced the judgment and operated as "halo."

Symonds (2) lists five reasons for the influence of

(1) Knight and Franzen -"Pitfalls in Rating Schemes"- Journal Ed. Psych. - Vol. 15:204.

(2) Symonds, Percival M.-"Notes on Rating" Journal of Applied Psychology-Vol. 9:188-195.

"halo" in ratings.

a-the trait or habit is not easily observed.

b-trait or habit is not commonly observed and thought about.

c-the trait or habit is not easily defined.

d-the trait or habit involves reactions with other people, rather than more personal behavior.

e-The trait or habit is one with high moral importance in its usual connotation.

We are now faced with the problem; how can the influence of "halo" be minimized?

One way to minimize "halo" in judgments is by care in the construction of the scale. Hughes (1) lists three principles in the construction of the scale which, if followed, will minimize the effect of "halo".

1. Unity of Definition--It is essential that every term in the rating scale be defined as unambiguously as possible. Contents of a given term must not allow wide range of interpretation of its meaning.

2. Behavioristic Definition--Every item in the scale should be defined in terms of behavior. Habits and qualities can only be judged on the basis of their outward manifestation. Inferences concerning subjective qualities can only be made on the basis of tangible

(1) Hughes, W. Hardin-" General Principles and Results of Rating-Trait Characteristics" Journal of Educational Method- 4:1924-25.

objective reactions to given situations.

5. Relation to situation--Personal qualities or habits are manifest only in appropriate situations. If there is no opportunity for exercise of the habit in question, the rating of the habit is worthless.

For example, suppose we wish to measure the neatness of a teacher. We may build a scale of this sort.

Neatness	Ex.	V.G.	G.	P.	V.P.
1. Personal appearance					
2. Habits of work					

The rator expresses his judgment by placing a check mark in one of the five columns. One of the factors which will probably operate to govern his choice of one of the five columns will be his like or dislike for the teacher.

However, if we construct a scale as follows:

	Ex.	V.G.	G.	P.	V.P.
1. Does she place pencils, papers and other teaching tools carefully away in a drawer at the close of the day?					

We are recording our judgment of specific activity and there is less liability of the same degree of "halo" which was operative in the first type of scale.

We may go still further and help make the judgment more concrete:

Meticulous. Each paper and each pencil in proper place in drawer.	Places tools away neatly in drawers.	Places tools hurriedly in drawer. Drawer shows results of haste.	Throws tools hurriedly into drawer when she thinks of it.	Leaves papers and pencils scattered over her desk.

By asking the rator which type of activity most nearly fits the teacher, we are minimizing the influence of "halo" still more.

Another way of minimizing the influence of "halo" in single ratings is to provide a common mental background for a group of rators. This may be done in conference before the ratings are made, or better yet, if the persons using the scale could share in the construction of it, in some such way as the Army Rating Scales were built, a common mental background could easily be established. At least, the rator could be shown what to look for in the way of qualities or habits.

This leads us to the third suggestion,--the competency of rators. The qualifications of a competent rator have been indicated on the preceding pages. Rugg (1) found the average deviation for competent rators less than any other group.

(1) "Is the Rating of Human Character Practicable" Journal Ed. Psychology-Nov-Dec 1921 - Jan, Feb. 1922

Another suggestion for minimizing the influence of "halo" is given by Hughes (2) as: Freedom from Emergency. Other things being equal, ratings are more reliable when not affected by emergency. If there is no immediate need for the information, the rating will be more free from bias. If a rating is for the purpose of recommending a man for a position and the rator knows he need the position, his judgment is apt to be influenced by his knowledge of the need.

A fifth suggestion, also listed by Hughes, is in the handling of the scale.

In general, ratings are more reliable when made on a single quality at a time for an entire group than when made on all qualities for an individual. This suggests that the influence of a judgment on one quality is apt to be felt in the judgments of other qualities. Therefore, it is necessary as far as possible to dismiss all other qualities and habits from mind and concentrate on the single quality to be observed.

d. The lack of knowledge of the subject on the part of the rator is another evidence of the necessity of competent rators. This might be called by some an advantage due to the limiting of the influence of "halo",

(2) "General Principles and Results of Rating"-Journal Ed. Meth.- Vol. 4:1924.

but unless a rator knows a subject well enough, there will be several manifestations of observable qualities which the rator will overlook.

However, the opposite of this is true. Knowing the subject too well has a bad effect on rating, particularly in the operation of 'halo' on the judgment. Knight(1) points out that, 'the factor of acquaintance operates to make ratings more lenient, i.e., increases the over rating, and to make ratings less critical and less analytical, i.e., increases the influence of the 'halo' of the general estimate'.

At first glance, the problems involved by personal factors in judging, seem to present unsurmountable obstacles. Many times, for this reason, the rating scale is considered of no value as an instrument. How can we obtain reliable ratings on a given subject?

Freyd (2) lists seven statistical criteria for increasing the reliability of the results of rating. This is perhaps the most complete list although Hugg (3), in his experiences with the Army Rating Scale mentions most

(1) Knight, F. B.-"Effect of the Acquaintance Factor upon Personal Judgments"-Journal of Ed. Psych. Vol. 14 - p. 142

(2) Freyd, Max-'The Graphic Rating Scale'-Journal of Educational Psychology - Vol. 14: 1923 pp. 83-102

(3) Hugg, H.O.-' Is the Rating of Human Character Practicable?'-Journal Ed. Psychology - Nov. 1921 pp. 425-438; Dec. 1921-pp 485-501; Jan 1922 pp. 30-42; Feb. 1922- p 81-93

of the above, either directly or indirectly, while Watson (1) explains the principle of correlation as a means of increasing reliability of measure.

A. A method of increasing the reliability of measures used by Rugg (2) and Folsom (3) is by correlating these with I.Q. scores or other objective measures. Rugg attempted to correlate the scale scores of individuals and I.Q. scores of the same individuals. Folsom attempted to correlate classmates judgments of 166 men on eleven character traits in a small college, professor's judgments on four traits, performance record in social popularity and athletics, results of physical examination and scores of responses to a two-series of advertisements test.

Rugg found in his experiment with the Army Scale no correlation between the scale score of an individual and the I.Q. score. Freyd (4) came to the conclusion, "that it is next to impossible to make statistical comparisons between kinds of rating scales. It seems likely

(1) Watson, J.G.-"Experimentation and Measurement in Religious Education."-Association Press-1927 - pp 44-45.

(2) Rugg, H.O.-"Is the Rating of Human Character Practicable?" Jour. Ed. Psych.-Nov. 1921-pp.425-38; Dec 1921:485-501 Jan. 1922-pp. 30-42; Feb. 1922:81-93

(3) Folsom, Jos. K.-"A Statistical Study of Character" Pedagogical Seminary & Jour. Genetic Psych.-Vol. 24:393-12

(4) Freyd, Max-"A Graphic Rating Scale for Teachers" Jour. Ed. Research.-Vol. 8-p. 433.

that ratings could be evaluated by correlating them with some variable, such as an objective measure of the particular trait upon which the ratings are made. But ratings are ultimates in Psychology and cannot be evaluated with reference to a known criterion. If such a criterion were available, there would be no need to make ratings."

B. Another method of increasing reliability is to compare the ratings on the same subjects by the same judges for different months. This is rarely a safeguard against the influence of "halo". The judgment of the rator is as apt to be influenced by "halo" the second or third month as it is the first.

C. Another method is the comparison of the ratings on the same men by different judges. This is perhaps the best method yet mentioned. The average of three independent ratings is apt to be a more accurate measure of the existence of a trait than any single rating. Rugg (1) says: "Assuming qualified rators, the reliability of a judgment increased directly with the square of the number of judgments. If the probable error (P. E.) on a single judgment is 0.6745; of two judgments it is 0.476; of three judgments it is 0.389; and of four judgments it is 0.345.

(1) Rugg, H.O.-"Is the Rating of Human Character Practicable" Journal Ed. Psy. - Feb. 1922-P 1-63

D. Another index of reliability is the normality of distribution. This means that the scores for a large number of individuals selected at random ought to follow the normal Probability Curve, (1) which on a five point scale is indicated roughly by these central percentages: 7% - 24% - 38% - 24% - 7%. Stated in another way, we might say that in a large number of cases the total number of individuals would fall into their proper fifth of the scale as follows; 7% lowest and 7% highest, 38% in the middle; 24% between the highest and middle and 24% between the lowest and middle. If this is not the case, it is evident that certain portions of the scale are being neglected by raters, or else steps on the scale are not of equal value. Donnelly (2) found in the preliminary construction of part III of his scale that the few items scaled at the central portion of the scale caused the scale to break in two parts at the middle. It is pointed immediately to the need of revision of the scale in order to get more reliable scores or results.

E. The fifth index of reliability is the spread of

(1) Garrett, Henry E.-"Statistics in Psychology and Education-pp 74-116; longmans, Given & Co., New York 1926

(2) Donnelly, M.I.-"Measuring Certain Aspects of Faith In God as Found in Boys and Girls Fifteen, Sixteen, and Seventeen Years of Age. - Westminster Press 1931-P 46

the distribution. Sufficient discrimination between qualities is essential in order to figure correlation coefficients and to distinguish between one individual and another. Too great a spread or range (1) increased the error of any single rating.

F. The sixth criterion for increasing the reliability of the score is the precluding of any possibility of 'halo' in the construction of the scale. This has been considered in the preceding pages.

From the above discussion it is evident that there are difficulties arising due tothe personal elements involved in the process of judging. Nevertheless, we believe the rating scale to be a reliable instrument providing:

a. The final score is the average of three independent ratings.

b. The scale has been carefully constructed.

c. The three rators are competent.

(1) Garrett, Henry E.- Statistics in Psychology and Education-P.17; Longmans Green & Co., New York 1926

CHAPTER IV.

THE RATING SCALE IN RELIGIOUS EDUCATION.

A. - The Rating Scale - A supervisory instrument.

In the preceding chapter the rating scale was discussed as an instrument of measurement. In this chapter we will attempt to show its use in the field of Religious Education as a supervisory instrument. This leads us to our first question. What do we mean by supervision?

Religious leaders, including many so-called religious educators, use the terms 'administration' and 'supervision' interchangeably. Supervision is confused with management, administration and executive work although administration, in the broadest sense, includes supervision.

Supervision, in the broadest sense, includes all that has to do with the teaching practice, i.e., the classification of pupils, teacher training, adjustment of class periods etc., while in the narrow sense, it means the improvement of those who actually teach.

Thus the purpose of supervision is two-fold:

(a) The attainment of increased skill on the part of the teacher.

(b) The efficient education of pupils.

In the light of this two-fold purpose, the supervisor will proceed, through conferences with groups and individuals, observation of actual practices, and demon-

stration work of various kinds, toward the goal of better teaching and better educational opportunities for pupils. How will he be able to discern actual progress? He must have some method of measurement which will show relative improvement of teaching procedure and relative progress in the lives of the learners. One method is the use of rating scales. Thus, the rating scale will be used by the supervisor in the rating of teachers, the rating of pupils, the rating of the program.

However this is not as simple as it would seem. As has been pointed out in the preceding chapter, the rating scale, to have value, assumes time for at least three independent ratings, well constructed scales and competent rators. This furnishes problems in the field of religious education.

The first problem is lack of adequately trained supervisors. The field of religious education is not yet sufficiently developed to provide a large number of religious educators who are competent supervisors. The large majority of present so-called religious educators have not had the background and training to enable them to build and use rating scales. Nor have the churches yet developed the consciousness of the need for this type of activity.

If we could at the present time assume that every

director of religious education was a competent supervisor, the problem would be lessened but would not be met, because it is next to impossible for a director adequately to supervise every phase of a modern program of religious education. Competent supervisors must also be trained from among volunteer workers to share in the program of supervision. However, we believe this to be a temporary condition. Books are appearing in the field, (1) courses are being offered in schools of religious education, courses are offered as part of the Standard Leadership Training Curriculum, (2) conferences are constantly being held (3) and in time the need which we feel very keenly now will be sufficiently met to make an adequate supervisory program possible.

A second problem which arises in the program of rating as a supervisory function, is the fact that teachers are not competent rators, i.e., competent from the standpoint of understanding the idea of measurement, familiarity with rating scales, familiarity with the qualities under observation, knowing pupils both in and out of the Church

(1) Chave, E.J.-"Supervision in Religious Education" University of Chicago Press 1931
McKibben, Frank M.-"Improving Religious Education Through Supervision"-Methodist Book Concern 1931

(2) International Council of Religious Education-Bulletin 503.

(3) Professional Advisory Sections of International Council of Ril. Ed.

School, and practice in the use of rating scales. This also, we believe to be a temporary condition. Teachers' meetings, training classes, and personal conferences between supervisor and teacher, will eventually help in meeting this need.

A third problem facing the supervisor is the fact that at present the time is inadequate in the so-called "Sunday-school" for a practical scheme of rating to be carried on. However, Vacation Church Schools, week-day schools of religion, club and society activities are being introduced and added to the traditional program, thus meeting the problem of time.

This points to a fourth need. Activities such as the above are too often separate organizations and entities, utilizing different leadership and developing separate and independent programs. When this is true, the problem of providing time for adequate rating is not met, because of different leadership and programs. Only as the total program is unified to allow the additional time for observation of activity by individual rators, will the rating scale become a valid supervisory instrument. The lack of correlation between activities and types of activities has given rise to a movement toward the unification of the total program. Many churches have unified their programs with very satisfactory results.

These problems, we believe, while at the present are evident in greater or lesser degree, will not necessarily militate against an adequate rating procedure as a part of the function of supervision, but can be eliminated by the training of directors of religious education, supervisors and teachers, and by building a program which utilizes vacation schools, societies, clubs, week-day schools, and by the unification of both organizations and programs.

B. - The Rating of Teachers

The rating scale for teachers might well try to evaluate two types of data: (a) personal qualities and habits which are desirable in the teacher of religion, and (b) skill in teaching. Scales are of two forms: (1) the self-rating scale and, (2) the scale to be used by the supervisor in judging personality qualities and skills of the teacher.

At the present time there is little use made of any type of evaluating instrument in the selection of church school teachers. A superintendent selects almost any one he can get to take a class, although in many cases the church board of religious education passes upon his selections. There is real need for ratings as the basis for recommending prospective teachers to the board of religious education for vacancies in the teaching staff of the church.

It is also necessary for the supervisor to know in rather detailed fashion the personal qualities and traits of teachers in service in order to aid him in counseling with and guiding individual teachers, and to aid in placing teachers in the age group for which they are best fitted.

Without the curve of salary increase or promotion in position which exists in secular education, but with a motive far stronger than either of these, i.e., belief in religion and conviction that the teaching of religion to boys and girls is the greatest thing in life, the supervisor has a real opportunity in assisting teachers to cultivate desirable personal habits in order to make their personalities more attractive to their pupils.

Given a vital religious conviction on the part of the teacher and a desire to improve his personal life, the self-rating scale offers opportunity for self-evaluation and points out the desirable qualities, thus acting as a spur to self-improvement.

Chave (1) suggests the first self-rating scale for the personal qualities of religious teachers. This is a five point scale of four divisions, i.e., personal qualities, leadership qualities, attitude toward work, and religious qualities. A sample of the scale follows:

(1) Chave, E.J.-"Supervision of Religious Education" "University of Chicago Press" 1931 page 316

CHARACTERISTICS MANIFEST IN RELIGIOUS EDUCATION	RATING				
	Best 10%	Next 20%	Middle 40%	Next 20%	Lowest 10%
	Ex	Good	Average	Poor	Bad
	1	2	3	4	5
Personal Qualities					
Attractive Personality - - - - -					
Common sense - - - - - - - - - -					
Industry - - - - - - - - - - -					
Perseverance - - - - - - - - - -					
Reliability - - - - - - - - - -					
Punctuality - - - - - - - - - -					
Originality - - - - - - - - - -					
Patience - - - - - - - - - - -					
Sympathy - - - - - - - - - - -					
Sincerity - - - - - - - - - - -					
Cheerfulness - - - - - - - - - -					
Self-control - - - - - - - - - -					
Adaptability - - - - - - - - - -					
Intellectuality - - - - - - - -					

This is not a good scale as it now stands because the statements of the qualities is ambiguous, they are not stated as specific habits, and they are not stated with reference to life situations. If our discussion of the elimination of halo in the construction of the scale is correct in chapter three, this scale would probably produce ratings which are influenced by halo. However, this is significant as a beginning and it can be made into a good self-rating scale. The same scale, revised as per the above suggestions, could be adapted to use as a diagnostic instrument for the supervisor, both in the selection of teachers and in the placing of teachers.

The self-rating scale is also a valuable supervisory aid in stimulating development in the function of teaching. As yet there are no scales of this sort available for the field of religious education as such, but Rugg's (1) self-rating scale for teachers in the secular field is a type which might well be utilized in the religious field. This scale appears in two forms. Form A is a three point scale of five divisions, i.e., (a) Skill in teaching, (b) Skill in the mechanics of managing a class, (c) Team work qualities, (d) Qualities of growth and keeping up to date, (e) personal and social qualities. Form B is a man-to-man comparison, of the same qualities as in form A. There are certain advantages in this type of scale.

1. There is little or no overlapping of qualities.
2. Concrete questions are asked in sentence form.
3. The teacher rates himself in simply one of three groups.
4. The teacher rates himself on the same form on which he is rated by the supervisor.

A sample of the scale follows:

Form A

II Skill in the mechanics of managing a class.

	Low	Average	High
To what extent— 1. Does the class work proceed smoothly (without artificial interruptions and transitions from one kind of discussion to another.)			

(1) Rugg, H.O.-"Self Improvement of Teachers Through Self-rating" Vol 20 Elementary School Journal pp 670-684.

	Low	Average	High
2. Do the pupils attend naturally and spontaneously to the work of the lesson.			
3. Does order or discipline inhere in the work (not maintained by compulsion or suppression).			
4. Is routine, as passing material, moving to the blackboard, etc., economically and systematically organized.			
5. Is material and equipment in the room effectively arranged.			
6. Does he pay attention to the details of heat, light and ventilation.			

Summary rating

Form B

II Skill in the Mechanics of Managing a class

Best Teacher	38	
Better than Average	30	
Poorer than Average	22	
Poorest teacher	14	
	6	

It is necessary to follow any attempt at self-rating by a conference between the supervisor and the teacher in order that a perfect understanding is effected. The supervisor suggests lines of study, activity, etc., which might help the teacher and together they work out solutions to the weaknesses apparent in the measurement.

While self-rating and self-evaluation are the bases which

provide the teacher with the necessary data for self-improvement, it is also necessary for the supervisor to take independent ratings from time to time in an effort to discover progress of the teacher toward the desired and mutually recognized ends. The instruments used in this process is an adaptation of the rating scale known as a check list or activity analysis. Barr and Burton (1) describe the activity analysis as "what does, might, could, should happen in a lesson. It is a statement, in as exact terms as possible, of small items of actually observable behavior on the part of the teacher or the pupil." A sample of a check list follows:

V Some Elements of Weakness often Present in Lessons

Some of the common faults observed in recitations are listed by way of suggesting the type of analysis that should be made.

1. The teacher utilized the wrong lesson activity: that is, she taught a development lesson when her class was ready for a practice lesson or perhaps for a review lesson; or she gave a review of a topic when practice or drill was needed.

2. Memory was the chief ability required to succeed in the recitation.

3. The teacher did the organizing for the pupils.

4. One or two pupils did all the talking.

5. The subject matter was treated as if all the facts were of equal importance.

6. In trying to meet individual needs the teacher neglected the majority of the class.

7. Only one pupil at a time responded in a test lesson where by a change of device all could have responded;

or, by substituting a socialized activity for a question and answer type of lesson, a much more general participation could have been secured.

8. The teacher passed judgment as to the concreteness of statements, thus robbing the pupils of the privilege of judging, and hence, making learning less sure.

9. The real subject matter of the lesson lay outside of the text and was not touched upon.

10. The pupils were not the actors; they were chiefly acted upon by the teacher.

11. There was little or no problem solving, hence, little or no thinking.

12. Up-to-date methods of teaching were seemingly unknown to the teacher, such as diagnosis before drill, individual practice to meet individual needs, and so forth.

13. The methods of study of the pupils were not taken into account nor fostered by the lesson plan.

14. No questions were asked by these pupils. Pupils who are really thinking ask questions.

Any method of scoring an activity analysis which is easiest and most understandable to the observer may be used. The data gathered by the check list is used by the supervisor as the basis of a conference with the teacher in pointing out strength and weakness in the function of teaching. Data gathered in this way is also used to objectify ratings on the teacher observed.

Thus we see that the rating of teachers has value in the supervisory function chiefly as the basis of discovering need for improvement both in the personal habits and skills in teaching, and as a spur toward achieving that improvement. Therefore, unless

(1) Barr and Burton-"The Supervision of Instruction" Appleton and Company 1926 Page 116.

the teacher is motivated by a desire for improvement and is anxious and willing to take suggestions from the supervisor, the rating scale has value only as an instrument which may be used in gathering information. The importance of the teacher's ability and willingness to take suggestion leads Wagner (1) to suggest a new unit of measurement in scoring rating scales. Instead of marking good, poor, fair, etc., he suggests a measure which may be defined as the teacher's need for suggestion:

> A score of 5-means-does not need suggestion, always asks for suggestions of other teachers or supervisor.
>
> 4- Teacher rarely needs suggestions. Often asks for them and adopts the ones needed.
>
> 3- Needs suggestions very often but rarely or never asks for them and infrequently uses those offered.
>
> 2- Cannot get along without suggestions, rarely succeeds in using them.
>
> 1- Can do nothing without suggestions or with suggestions.

However, the supervisor's function is not confined solely with teachers in service, but with prospective teachers who are in training as well. Senska (2) has formulated a "Detailed Score-Card for Rating Student Teachers" which is made up of the following divisions with appropriate descriptive subdivisions; Character, Scholarship, Teaching skill, Daily preparation, Discipline, Attitude, Interest in pupil activities, Classroom man-

(1) Wagner, C.A.-"The Construction of a Teacher Rating Scale" Elementary School Journal Vol 21, pp. 361-366.

ment, Personality and Health. A sample of the scale follows:

Characteristics	points	A Grade	C Grade	F Grade
Character	Standards of action	Lives closely to highest ideals	Working toward ideals	Violates standard of social group
	Service	Helpful in best way at right time	Helps when asked	Self first
	Fair mindedness	Can judge truly even when it hurts	Conclusions drawn from data at hand	Subject to prejudice
	Uprightness	Glad to be thought to be just what she is	Puts best foot forward	Deceives for the sake of supposed advantage
	Mastery of self	The spirit controls the flesh	Good purposes but not always able to carry out	Gives in to desires and whims
	Progressiveness	Alert to the best in the new. Ready to change for the better	Goes with the crowd	Extremes appeal or too set to change

The student teacher works under the direction of a critic teacher. During the first third of each term both critic and supervisor are expected to grade every student teacher on each of the ten classifications of the scale. These grades are used as a starting point for improvement with supervisor, critic and student working toward that end. An estimate is again taken at

the end of the year and the final grade given.

An adaptation of this plan to the field of religious education would be both desirable and feasible. Ratings by critic, supervisor and a self-rating of the student teacher would provide the basis for determining a rather valid estimate of progress.

Thus the rating scale becomes a very valuable tool for the supervisor in religious education as an instrument in the selection of likely prospects as teachers, as the basis for improvement of teachers in service, and as an instrument for guiding student teachers.

C. The Rating of Pupils

The supervisor's interest, in addition to the problems of improving teachers in the teaching function, must also be very evident in the actual progress of the learner. Pupil rating scales will be of two types: (a) self-rating scales for pupils and, (b) scales for teachers and leaders which will help them in discovering progress made by pupils in the development of character, which we conceive to be the basic purpose of religious education. Neumann (1) presents a word of caution in this regard; "Character rating is different from rating academic achievement or needs. It is possible to be more objective about one's proficiency in spelling than about the purity, let us say, of one's unselfishness," and yet observable, objective data must

(1) Neumann, Henry-"Some Doubts About Character Measuring" Ril Ed. Vol 5:620-626

be gathered if we are to make valid inferences as to the existence or non-existence of desirable, intangible qualities.

The observations which have been made regarding self-rating for teachers and the urge toward better personal life and better skills in teaching, might just as easily be made for the self-rating and self-evaluation of personal character habits on the part of pupils. Neumann (1) points out, "Charts in which the pupils do their own rating of themselves have indeed a certain usefulness. Pupils are more likely to improve when they are taught to look for more truthfulness, industry, cleanliness, courtesy, rather than for character in general or for some one trait which happens to be especially interesting to the teacher. When emphasis is put on the constant need to improve, the pupil gets something of the spur there is in playing an interesting game. A boy charts a graph for himself, to indicate, for instance, how steadily or otherwise he manifests some ten or a dozen desirable traits. He observes that his profile is a very jagged looking affair. His teacher lets him see another pupil's in which the line is less rocky. Next month the boy tries to have his own line straighter. Many, no doubt, are benefited by such a periodic look at themselves in the moral mirror."

Thus with the aim of self-realization or personal development constantly before the group, the self-rating scale for pupils becomes not only desirable but necessary. The Presbyterian

(1) Neumann, Henry-"Some Doubts About Character Measuring" Ril Ed. Vol 5:620-626

Board of Christian Education (1) has constructed the first pupil's self-rating scale of this sort and has embodied it in a program which is designed to stimulate just such a desire for individual development. This is a six point scale of one hundred and fifty-seven concrete questions concerning life habits which are classified under the following headings:

I. Getting and Caring for Things that are Valuable.
II. Getting an Education.
III. Making and Keeping Friends.
IV. Keeping Faith with Sex and Family Ideals.
V. Making, Building, and Inventing things.
VI. Appreciating things that are Beautiful.
VII. Seeking Safety and Peace.
VIII. Showing Reverence for God and for Religion.

A sample of the scale follows:

Give yourself a fair and square rating for each checking point.

VIII. Showing Reverence for God and for Religion.

			1. Am I reverent in my behavior and attentive when I go to church?			
			2. Am I respectful toward people whose church and religion are different from mine?			
			3. Do I cheerfully do my part in the activities of my church and Church School?			
			4. Am I regular and prompt in my religious duties?			
			5. Am I learning what good it does to pray?			
			6. Am I coming to know why it is that people think so much of the Bible?			
			7. Do I let God help me with my difficulties?			
			8. Am I learning how to join with other people in a service of worship?			

(1) Religious Education Vol 5 pp. 622

			9. Do I enjoy doing what is right just because it is right in God's sight?					
			10. If I find out that anyone is in need, do I consider it a religious duty to try to help him out?					
			11. Do I show respect for my religious leaders?					
			12. Do I keep from swearing, always being reverent in using the words, "God" and "Jesus Christ"?					
			13. Am I careful not to permit others to neglect their religious duties?					
			14. Am I doing what I can to spread Christian friendliness throughout the world?					

The club program consists of a series of suggested activities, in which one or more members of the club can engage, which are designed to put into practice the habits or qualities described by the questions on the scale. Thus if a boy finds a very definite weakness which he wishes to correct, he chooses the type of activity in which he can practice doing effectively the type of thing under guidance which will help to remedy weaknesses which are apparent to him.

In this way the self-rating scale for pupils, provided the pupil is properly motivated with the desire for self-improvement, becomes a diagnostic instrument which helps in the discovery of individual needs, and an evaluating instrument which helps in evaluating progress toward desired ends.

Rugg (1) presents a self-rating scale for rating pupils' dynamic qualities which is divided into the following divisions: (a) Ability to learn-to assimilate new ideas, (b) Qualities of industry and attitude toward school work, (c) Qualities of leadership, (d) Team-work qualities, and (e) Personal and social

(1) Rugg, H.O. "Rating Scale for Pupils' Dynamic Qualities" School Review Vol 28:pp. 744-5

qualities. This scale is identical in form to the scale for teachers which has been described in the preceding pages. It may be used either as a self-rating scale or as an instrument in the hands of teacher or leader.

While self-rating of students is highly desirable from the standpoint of the learning process, teachers, leaders, and supervisors will wish to take ratings on pupils from time to time in order to discover the effectiveness of the teaching function, and dominant needs to be met by the program.

Cornell, Coxe and Orleans, of the Educational Measurements Bureau, New York State Department of Education, (1) have devised a grphic scale for rating school habits as an instrument to be used by teachers. This scale deals with the following qualities; Attention, Neatness, Honesty, Interest, Initiative, Ambition, Persistance, Reliability and Stability.

A sample of the scale follows:

Attention

Extreme inability to give attention to task	Usually pays attention can be distracted	Always pays very close attention while studying or during class periods.

Neatness

(1) Published by World Book Company, Yonkers-on-the Hudson

The teacher checks the student's position on the line which corresponds, in his judgment, with the pupil's status in regard to the quality observed.

A very pretentious attempt to measure individual growth in religious education is presented by the International Council of Religious Education (1). This is a five point scale of eleven areas of experience in which youth lives. These areas of experience are as follows: Health activities, Educational Activities, Economic activities, Vocational preparation, Citizenship, Recreation, Sex-parenthood-and family ideals, Personal Group Life, Friendship, Aesthetic Interests, Specialized religious activities.

The scale is constructed for the average Church leader. Each of these areas of experience is further defined with appropriate sub-headings, and the descriptions of the five points of the scale make it useable as the following sample will show.

Areas of Experience in which Youth lives.	Description of the five points of attainment.				
	Negative			Positive	
	1 Bad	2 Poor	3 Medium	4 Good	5 Excellent
3. Economic Activities (1) Attitude toward money	Regards it as a means of selfish enjoyment	Ignorantly or indifferent to right attitude in money matters	Desirable att't de when no personal sacrifice is involved	Unselfish and generous	Creates Stewardship ideals in others

(1) Pamphlet #3-Christian Quest Materials "How to Study Individual Growth" 1937

(2) Use of money	Wastes, dishonest no effort to earn	In debt no accounts	Irregular accounts Balances welcome	Terms. Spends by budget. Saves. Gives.	All money used on Stewardship basis of life

The teacher or leader chooses the description which most nearly fits the pupil under observation for each area of experience, and checks it, thus discovering on which of the five levels from best to worst, the pupil lives.

Truly consecrated teachers and leaders in religious education catch a new vision of their function, when personal habits or qualities of their pupils are revealed to them by the use of rating scales. Pupils also catch a new vision of life when they discover aspects of their lives which need improvement. Proper use of rating scales in the analysis of personal qualities, may provide the starting point for a truly democratic process of education, i.e., teacher and pupil discovering together and sharing the best in life.

D. Rating The Program.

The supervisor's responsibility is not confined solely to the improvement of teachers and teaching and results in the lives of pupils. He is also concerned with the various factors which make this possible. A teacher may be a very good teacher, but, without adequate facilities with which to work, best results are not obtained. Equipment, housing facilities etc., are important factors in the whole educative process.

The supervisor in religious education has responsibility also for other elements than those which we commonly associate with teaching and classroom activity. The program of religious education is composed of the elements of worship, study, recreation and social life, and service. Study and service are commonly associated with class activities, and teacher-pupil relationship. Worship and recreation are usually group activities in which several classes participate. They are no less a part of the educative process, however, because of their influence on the lives of the participants. It is therefore necessary for the supervisor, in addition to the improvement of teaching to guide the worship and social activities of the church into the most efficient constructive channels possible. Development or progress of this sort

implies the need of some sort of instrument with which to evaluate the status of the program at any given time, - the needs to be met, and improvement made.

The rating scale is peculiarly adapted to uses of this sort and has already been used for this purpose in religious education, particularly in score-card form.

For the sake of further examination we might classify score-cards as; (a) scales for measuring the whole program and (b) scales for measuring particular phases of the program.

1.- Measuring the Whole Program.

If we are committed to the seven objectives listed in the first chapter as the desired goals or ends toward which we must strive in the construction of a program of religious education; and if we are committed to the principle that education is a process of gradual, unfolding development, the elements which must be built into the program must be varied in their nature, in order to provide the conditions in which the sense of a personal relationship to God, an appreciation and understanding of the teachings of Jesus, the development of

a Christ - like Character, a passion and concern for the welfare of society, and a new vision concerning the function and the mission of the church, can be made possible. We are very much concerned that the elements of the programs, i.e., worship, study, service, social and recreational life, and personal experience in religion and the church will be provided, and that each element will be given proper emphasis and each aid in evaluating the total program.

These instruments are in score-card form, and deal with the general classifications of Curriculum, Leadership, Organization and Administration, and Housing and Equipment.

The supervisor judges his program against a standard, Scoring is done on the basis of 1000 points. A sample of Standard A follows:

I Curriculum	Perfect Score	School Score	Total
1. Worship	70		
2. Service	65		
3. Study	70		
4. Social & Recreation life	55		
5. Personal Experience in religion and the church	65		
Total for Curriculum	325		

Detailed description of the standard are made to help the rator as the basis for judgment, i.e.,

5. Personal Experience in Religion and the Church

Religious education should lead each pupil to a personal faith in God, acceptance of Jesus Christ

and his way of life, and membership in the Church. Membership and participation in the life and work of the Church is not only an expression of loyalty to the cause of Christ but a primary means of growth in Christian living.

(1) Is effort made to help each pupil develop an intelligent faith in God and an increasing devotion to Jesus Christ and his way of Life?

(2) Is effort made to lead those who are intelligently and spiritually prepared therefore to personal commitment to Jesus Christ and formal reception into the church?

(3) Is provision made for training in the meaning and duties of church membership?

(4) Is special effort made to deepen the interest and increase the activity of the new church members after they have been received?

Standard B, Standard for the Vacation School, Standard for the Week-day Schools, are all similar to the above. Departmental standards such as Standard for the Primary Department (1) are also instruments for measuring the total program, even though the age group for which it is used is restricted.

(1) International Council of Religious Education Chicago, Ill.

Assuming that the supervisor has well defined goals for the program he hopes to build, score cards of this sort help in three ways:

(a) as a diagnostic instrument to discover status quo.

(b) as an evaluating instrument to discern progress toward well defined goals.

(c) as a publicity instrument in aiding the discovery of basic needs, and in defining them to a church school board, as a spur to further progress.

2. Measuring parts of the program.

Suppose the supervisor was to discover from an evaluation of the total program that the study element was too greatly stressed and that worship, recreation, etc., were, by comparison, weak, his first question upon the discovery of a weakness in a given part of the program, would be, where is this part weak and how do I start in to correct it? Hence, in addition to measuring the total program, the supervisor must measure, from time to time, elements in or parts of the program. This is practically what we are doing when we utilize a check list in the observation of a teacher in action, or a

general rating of teacher or pupil activity of a specialized type.

(1) The Standard for City Church Plants, which has been described in chapter three, in attempting to discover the useability of city church plants, is another effort to measure a specific factor in the program.

Thus far, however, there are no scales available for measuring worship activities, social and recreational elements, service activities, and church contacts and experiences, which the supervisor might utilize. If these elements are to be measured, scales for this purpose must be constructed by the supervisor himself. However, a committee of ten members of a class on surveys and

(1) Published by the Interchurch World movement.

Measurement in Religious Education at Boston University attempted to build a score card for rating the quality of worship, a sample of which appears herewith:

The main factors this experiment purposed to measure were:

I Room and Equipment
II Worship Program--Materials
III Worship Program--Conduct
IV Leadership
V Attitude of the Worshippers

Samples of scoring and descriptive definitions for scoring follow.

	Perfect Score	Service Under Observation
III Worship program,---its conduct Total 200		
A. Freedom from distracting elements	30	
B. Teachers, officers and pianist in close cooperation	50	
C. Discipline by interest	90	
D. Length of service	30	
TOTAL		

The descriptive standard is as follows:

III Worship program,---its conduct.

A. Freedom from distracting elements
a. Doors should be closed to tardy pupils until certain point in the service.

b. Visitors should be unobtrusive

B. Teachers, officers and pianist in co-operation with the leader
   a. Responses should be prompt
   b. Prayers, solos, etc., should be in harmony with the theme
   c. Transition between parts should be smooth
   d. There should be a minimum of impromptu contributions.

C. Discipline by interest.
   a. Theme should be chosen with the pupil's interest in view
   b. Interest should be sustained by participation

This projected scale is still undergoing revision and may some day be available.

Experimentation of this sort will result eventually in providing adequate score cards and standards for the elements in the program which are not now being measured.

Thus we see that the score card or rating scale is a valid supervisory instrument in evaluating the total program of religious education as well as separate elements in the program.

CHAPTER V

CONCLUSIONS

If the rating scale is carefully constructed, if the rators are competent, and if the final rating is the average of three independent ratings by different rators, the rating scale is the most practicable measuring instrument available for use in the supervision of religious education.

CHAPTER VI

SUMMARY

SUMMARY

Despite the fact that there is great need for refinement in the process of judging the degree and the nature of the intangible qualities which make for success in business, industry, education, and similar walks of life, no where is the need more evident than in the field of religious education. This is due to the fact that the desired outcome of religious education is a process of developing a Christ-like character which expresses itself in relationship to God, in relationship to society, and in relationship to the Church, and that the phenomena to be measured consist of knowledge, attitudes and motives, conduct and the inter-relationship of all these.

The measurement movement to date has been chiefly concerned with tests, although significant progress has been made in the development and use of rating scales. Attempts at measurement in religious education have been made largely of test type in which

biblical knowledge, religious ideas, ethical discrimination, attitudes, and conduct tests have received major attention.

However, the assumptions that intangible qualities exist; that they can be so defined as to be easily discovered; that they are present in greater degree in some persons than in others; that they can be inferred from behavior; that difference in the degree of their manifestation is discernible; that convenient scoring method will make for concreteness; that accuracy of judgment is possible for carefully defined types of behavior; and that in the cultivation of desirable qualities, certain instruments are necessary for periodic evaluation of progress have been responsible for the development of four types of rating scales, i.e., the simple scale, the graphic scale, the man-to-man comparison scale and the score card. The rating scale has not been universally popular and definite opposition has expressed itself in the question of the existence of traits, as non-statistical criticism in the shape of problems of administration,

and statistical criticism which doubts reliability of measures obtained in this fashion. However, well constructed scales and the average of three independent ratings from three competent judges, produce a degree of reliability which justifies the use of the rating scale.

Conditions in the field of religious education at present, except in a few instances, hardly warrant the use of the rating scale due to lack of competent rators, well constructed scales and time for three independent judgment. However, this is a temporary condition and with a better trained class of directors of religious education, the rating scale will become the chief supervisory instrument for the measurement of teachers and teaching, conduct of pupils, and the effectiveness of the program.

BIBLIOGRAPHY

Bibliography

Books

Barr and Burton "The Supervision of Instruction" D. Appleton & Co. 1926

Bower, W. C. "Religious Education in the Modern Church" Bethany Press. 1929

Coe, George A. "What is Christian Education?" Scribners 1929.

Chave, E. J. "Supervision in Religious Education" University of Chicago Press 1931.

Donnelly, M. I. "Measuring Certain Aspects of God, as Found in Boys and Girls Fifteen, Sixteen, and Seventeen Years of Age." Westminster Press 1931.

Fiske, G. W. "Purpose In Teaching Religion" Abbington Press 1929.

Garrett, Henry, E. "Statistics in Psychology and Education" Longmans Green and Co. 1926.

Hartshorne and May "Studies in the Organization of Character" MacMillan 1930.

Kyte, George C. "How to Supervise" Chs. 13 & 14 Houghton Mifflin Co. 1930.

Lotz and Crawford "Studies in Religious Education" Cokesbury Press 1931 Chs. 6-8-19-20.

McKibben, Frank M. "Improving Religious Education Through Supervision" Methodist Book Concern 1931.

Myers, A. J. W. "Teaching Religion" Westminster Press 1930.

Thurstone and Chave "The Measurement of Altitude" University of Chicago. Press 1929.

Veith, Paul H. "Objectives in Religious Education" Harpers 1930.

Watson, Goodwin B. "Experimentation and Measurement in Religious Education" Association Press 1927.

Articles

Boyce, Arthur Clifton "Methods for Measuring Teachers' Efficiency" 14th Yearbook of the National Society for the Study of Education Part II.

Brandenburg G. C. and Remmers, H. H. "Rating Scales for Instructors" Educational Administration and Supervision Vol 13; 399-406.

Bralken, John L. "The Duluth System for Rating Teachers" Elementary School Journal. Vol.23; 110-119.

Brandenburg, George C. "Analyzing Personality" Journal of Applied Psychology June 1925: 139-155 Sept. 1925: 281-292.

Breslich, Ernest R. Gray, Wm. S., Pieper, Charles J. and Reavis, Wm. C. "The Supervision and Administration of Practice-Teaching" Educational Administration and Supervision Vol 11: 1-12 Jan. 1925.

Brown and Shelmadine "A Critical Study in the Objective Measurement of Character" Journal of Educational Research Nov. 1928: 290-296.

Brown, Edwin J. "A Character-Conduct Rating Scale" Education Vol 50: 369-379.

Chassell, Clara F. "The Army Rating Scale Method in the Kindergarten" Journal of Educational Psychology Vol 15: 43-52 1924.

Cleeton and Knight "Validity of Character Judgments Based on External Criteria" Journal of Applied Psychology Vol 8: 215-231 1924.

Connor, William L. "A New Method of Rating Teachers" Journal of Educational Research Vol I p.p. 338-358.

Folsom, Joseph K. "A Statistical Study of Character" Pedagogical Seminary and Journal of Genetic Psychology Vol 24: 399-437 1917.

Freyd, Max "The Graphic Rating Scale" Journal of Educational Psychology Vol 14: 83-102 1923.

"A Graphic Rating Scale for Teachers" Journal of Educational Research Vol 8: 433-439 1923.

Furst, Clyde "A Simple Literal Personal Rating Scale" Educational Administration and Supervision November 1922 p 463.

Giles, J. T. "A Recitation Score-Card and Standards" Elementary School Journal Vol 23: 25-36

Haggerty, M. E. "Recent Developments in Measuring Human Capacities" Journal of Educational Research April 1921. 241-253.

Hanna, Joseph V. "Variable Factors Encountered in the Rating of Students" School Science and Mathmetics Vol 25: 481-488 1925.

Hill, C. W. "The Efficiency Ratings of Teachers" Elementary School Journal Vol 21: 438-443.

Hughes, W. Hardin "General Principles and Results of Rating Trait Characteristics" Journal of Educational Method Vol 4: 1924-1925.

Knight, F. B. Franzen, R. H. "Pitfalls in Rating Schemes" Journal of Educational Psychology Vol 13: 204-213. 1922.

Lindsay, E. E. "Personal Judgments" Journal of Educational Psychology October 1921 413-415.

Neumann, Henry "Some Doubts About Character Measuring" Rel. Ed. Vol 5: 620-626.

Robertson, David Allan "Character Processes in Colleges and Universities " May 1930: 393-397.

Rugg, H. O. "Rating Scales for Pupils' Dynamic Qualities; Standardizing Methods of Juding Human Character" School Review Vol 28: 337-349.

"Is the Rating of Human Character Practicable?" Nov. 1921: 425-438; Dec. 1921: 485-501; Jan 1922: 30-42; Feb. 1922. 81-93.

"Rating Scale for High School Students" Journal of Educational Psychology Vol 12: 1921 p 431.

"Self-Improvement of Teachers Through Self-Rating" Elementary School Journal Vol 20: 670-684.

Senska, Nellie M. "A Detailed Score-Card for Grading Student Teachers" Educational Administration and Supervision" Vol 11: 199-201 March 1925.

Shuttleworth, Frank K. "New Method of Measuring Character Traits" School and Society Vol 19: 679-682 1924.

Spencer, P. R. "A High School Principal's Self-rating Card" School Review Vol 30: 268-273.

Lymonds, Percival M. "The Present Status of Character Measurement" Journal of Educational Psychology Vol 15: 484-498.

"Notes on Rating" Journal of Applied Psychology Vol 9: 188-195.

Touton, Frank C. "A Self-rating Score Card for Secondry School Principals" Journal of Educational Research Nov 1923: 335-345.

Wagner, C. A. "The Construction of a Teacher Rating Scale" Elementary School Journal Vol 21: 361-366.

Witty, Paul A. and Lehman, Harvey C. "The So-called General Character Test" Psychological Review Nov 1927: 401-413.

Yepsen, Lloyd N. "A Score Card of Personal Behavior" Journal of Applied Psychology Feb. 1928: 140-147.

BOSTON UNIVERSITY
1 1719 02575 1233